Quantifying the relative change in maintenance costs due to delayed maintenance actions on transportation infrastructure.

Zachary Hamida[*]and James-A. Goulet
Department of Civil, Geologic and Mining Engineering
Polytechnique Montreal, CANADA

April 19, 2024

Abstract

Identifying optimal maintenance policies for transportation infrastructure such as bridges, is a challenging task that requires taking into account many aspects relating to budget availability, resource allocation and traffic re-routing. In practice, it is difficult to accurately quantify all of the aforementioned factors; accordingly, it is equally difficult to obtain network-scale optimal maintenance policies. This paper presents an approach to evaluate the costs associated with deviations from optimal bridge-level maintenance policies, specifically focusing on delays in maintenance actions. Evaluating the cost of maintenance delays is performed using a reinforcement learning (RL) approach, that relies on a probabilistic deterioration model to describe the deterioration in the structural components. The RL framework provides estimates for the total expected discounted maintenance costs associated with each maintenance policy over time, allowing to compare maintenance policies where maintenance actions are delayed, against an optimal maintenance policy. The comparisons are performed by probabilistically quantifying the ratio of expected costs associated with each maintenance policy. This ratio represents the trade-offs between performing or delaying maintenance actions over time. Moreover, the proposed approach is scalable, making it applicable to bridges with numerous structural elements. Example of application using the proposed framework is demonstrated using inspection data from bridges in the Quebec province, Canada.

1 Introduction

Prioritizing bridges for maintenance is a challenging task, as these decisions are often conditional upon the availability of budget and resources required to perform the repairs, as well as the potential for managing traffic disruptions resulting from a partial or full closure of bridges [33, 7, 21, 40]. The aforementioned factors impose restrictions on taking optimal maintenance actions, and could cause delays in performing timely repairs [32, 7, 40]. Such delays can impact the future state of bridges, and can also incur higher repair costs, as the likelihood for major repairs or replacement becomes higher [7]. Therefore, it is important to have quantitive insights about the expected costs associated with delaying maintenance on bridges. Such costs represent an essential component in bridge life-cycle costs analyses (BLCCA).

Costs in BLCCA can be categorized into two types of costs, routine costs and extraordinary costs [16, 3]. Routine costs refer to the periodic spending on maintenance, cleaning and inspections, while extraordinary costs represent the expenditures when unforeseen and infrequent events occur, such as severe floods and earthquakes [16, 6]. Proactively minimizing routine costs serves as a preventive measure against the escalation of minor issues into substantial and financially burdensome problems. In addition, it fortifies preparedness to effectively cope and manage extraordinary costs when they arise. Therefore, infrastructure owners seek to minimize the total routine costs by identify the most economically efficient strategies.

Typically, existing methods for minimizing costs during the service-life of bridges rely on two interconnected phases, 1) optimizing the repair decisions on a bridge, and 2) schedule maintenance

[*] Corresponding author: blanche.laurent@polymtl.ca

decisions for a given set of bridges. The optimization of repair decisions on a bridge relies on several models, including deterioration models, Markov Decision Processes (MDPs), and reliability-based optimization [1, 5, 22]. In the context of condition-based maintenance (CBM), the combination of deterioration models and MDP-based frameworks are prevalent in the literature [2, 19, 39, 14]. Deterioration models provide estimates for the health states of the structural components within bridges based on inspection data. The health states refer to different levels of deterioration, ranging from perfect to poor condition. Decision-making frameworks take into account the health states of the bridge components as well as all possible actions that could alter them. A decision-making process using MDPs involves determining the optimal maintenance actions at each time step, by considering the current health state and potential future deterioration or improvement trajectories. As bridges typically have a large number of components, identifying an optimal maintenance policy becomes challenging due to the large number of state and action combinations in the MDP framework. Accordingly, recent maintenance planning frameworks have relied on deep reinforcement learning (DRL) for identifying optimal maintenance policies [2, 39, 37, 19, 38, 14]. DRL allows searching and identifying optimal maintenance policies by estimating and minimizing the total expected discounted costs associated with each pair of health state and maintenance action [34]. Applications of DRL has been shown to be effective in problems with a deterministic and large state-space, by learning policies that minimize the expected total costs associated with maintenance actions [2, 19, 39]. Nonetheless, for bridges with large number of components, the practical use of DRL methods is limited [14]. This is because the number of decisions at each time step is equivalent to the number of structural elements in the bridge, which makes it challenging to learn optimal maintenance policies and schedules over time. The complexities of maintenance planning on a single bridge are further compounded when multiple bridges are involved in the planning scope. Network-scale maintenance planning requires including additional decision factors (e.g., network connectivity and re-routing cost) to act as constraints on the planning problem [7, 33, 40, 9, 20]. In the presence of such constraints, it is challenging for decision-makers to perfectly follow the optimal maintenance policy required for each bridge in a network of bridges.

The primary objective in this paper is to address the following question: If a bridge is a candidate for maintenance intervention, and an optimal maintenance policy is already established, how would deviating from this optimal policy and delaying maintenance impact the total maintenance cost over time? To address this question, a probabilistic model is formulated for estimating the cost of maintenance delays at both the structural element level and the bridge level. The proposed framework relies on the capacity of RL to estimate the expected value for the current and future costs associated with each pair of maintenance action and deterioration state. The deterioration states in this context are estimated using state-space models which takes into account the inspectors uncertainty, in addition to providing estimates for the deterioration speed over time. The contributions in this paper can be summarized as follows: 1) the probabilistic quantification for the relative change in maintenance cost due to maintenance delays, 2) estimating the probability of replacement over time for a structural element given a maintenance policy, 3) an interpretable and scalable reinforcement learning approach for bridges with large number of structural elements, 4) validating the new approach using monitoring data from the network of bridges in the Quebec province, Canada.

The rest of the paper is structured as follows: firstly, the problem formulation is presented, followed by a background section to provide insights into the theory of sequential decision-making and reinforcement learning. Next, the methodology is introduced, which outlines the proposed approach for quantifying the impact of maintenance delays on the maintenance cost. Subsequently, we showcase an example application on a bridge using the developed method, followed by a discussion about the advantages and limitations, then a conclusion section.

2 Problem Formulation

A bridge $\mathcal{B}$ is composed of a number of structural elements e_p^j which are visually inspected by different inspectors $I_i \in \mathcal{I}$ over time. The health condition $\tilde{y}$ is represented on a continuous scale $\tilde{y} \in [l, u]$, where a structural element is considered in a perfect condition when $\tilde{y} = u$, while $\tilde{y} = l$ represents the worst condition. The presence of $\sim$ in $\tilde{y}$ implies that the variables are in a bounded space which correspond in this study to $[l, u] = [25, 100]$ [12].

Routine costs during the service-life of bridge $\mathcal{B}$ includes the cost of maintenance, inspection and other related costs, such as the cost of rerouting and time-value due to congestions resulting from maintenance/inspections [16]. Accordingly, the total routine costs $x_{r,t}$ at any time t can be described by,

$$x_{r,t} = \sum_{p=1}^{\mathtt{E}} r_p(s_t, a_t) + \sum_{p=1}^{\mathtt{E}} x_{i,p} + \epsilon_t, \tag{1}$$

where $\mathtt{E}$ is the number of structural elements in the bridge, $r_p(\cdot)$ is the cost of repairs associated with the p-th structural element, s_t, a_t represent the health state of the structural element and the maintenance action at time t, $x_{i,p}$ is the inspection cost for the p-th structural element, and ϵ_t corresponds to other costs associated with the maintenance or inspection operations. Under the assumption that the inspection schedule is fixed, minimizing the total routine costs over time can be formulated as a maintenance planning problem such that,

$$\mathcal{L}_c(\boldsymbol{a}_t) = \sum_{t=0}^{\infty} \gamma^t \left[\sum_{p=1}^{\mathtt{E}} r_p(s_t, a_t) + \epsilon_t \right], \tag{2}$$

where $\mathcal{L}_c(\cdot)$ is the loss function representing the cost of maintenance operations, $\boldsymbol{a}_t$ is a set of maintenance actions at time t for all the elements in the bridge, and $\gamma^t \in]0, 1[$ is the discount factor over time t [23]. Considering all costs defined in the domain $[0, -\infty]$, the optimal set of actions can be obtained using,

$$\boldsymbol{a}_t^* = \underset{a_t \in \mathcal{A}}{\arg\max} \quad \mathcal{L}_c(\boldsymbol{a}_t), \tag{3}$$

where $\boldsymbol{a}_t^*$ is a vector of optimal actions at time t and the set $\mathcal{A}$ is composed of $\mathcal{A} = \{a_0, a_1, a_2, a_3, a_4\}$, corresponding to, a_0: do nothing, a_1: routine maintenance, a_2: preventive maintenance, a_3: repair, and a_4: replace [25].

Solving the optimization problem defined above is intrinsically difficult because it is composed of two challenging problems, 1) identifying $\mathtt{E}$ optimal actions for the structural elements at each time t, and 2) grouping and scheduling the optimal maintenance actions over time. Moreover, deviating from the optimal vector of actions $\boldsymbol{a}_t^*$ requires learning a new set of optimal actions to be performed during the service-life of the bridge. The aforementioned challenges provide reasons to rely on reinforcement learning (RL) methods for maintenance planning. The decision making in a RL framework is based on an optimal mapping between the health states and maintenance actions, so that any deviation from the vector of optimal maintenance actions at time t, does not require re-learning the optimal set of actions $\boldsymbol{a}_{t+1}^*$ at time $t+1$. The next sections describe the theoretical foundations for sequential decision-making and reinforcement learning.

3 Background

3.1 Markov Decision Processes (MDP)

A MDP is an approach for modelling sequential decision-making, where taking the action $a \in \mathcal{A}$ enables a transition from state s_t to s_{t+1}. Each action a taken in a MDP model affects a return (e.g., positive feedback or negative feedback), where the return, represented by G_t, is the sum of rewards over time, starting from the immediate reward r_t at time t [31]. Accordingly, the return reflects the overall quality of a decision taken in the MDP model, and provides a mechanism for comparing different decision-making schemes. Those mechanisms are formalized in practice by the value function $V_\pi(s)$ and the action-value function $Q_\pi(s, a)$. The value function corresponds to the expected discounted total return for being in a state s, under policy π, which can be written as,

$$V_\pi(s_t) = \mathbb{E}_\pi\left[G_t|s_t\right] = \mathbb{E}_\pi\left[\sum_{i=0}^{\infty} \gamma^i r(s_{t+i}, a_{t+i})|s_t\right], \tag{4}$$

where $\mathbb{E}_\pi$ is the expected value while following the policy π, $r(\cdot)$ is the reward function which denote the expected value of the reward given the state s and the action a, as in $r(s_t, a_t) = \mathbb{E}[R_t|S_t = s, A_t = a]$. The discount factor $\gamma \in]0, 1[$ enables formulating and solving infinite planning horizon problems [23]. On the other hand, the action-value function $Q_\pi(s, a)$ represents the expected discounted total return for being in a state s and taking an action a under the policy π, such that,

$$Q_\pi(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_\pi \left[\sum_{i=1}^{\infty} \gamma^i r(s_{t+i}, a_{t+i}) | s_t, a_t \right]. \tag{5}$$

Based on Equation (5), the optimal decision-making policy π^* is the one maximizing the Q function,

$$Q^*(s_t, a_t) = \max_\pi Q_\pi(s_t, a_t),\ \forall s_t \in \mathcal{S}, a_t \in \mathcal{A}. \tag{6}$$

Identifying optimal policies within a MDP formulation can be done using different approaches, such as, reinforcement learning (RL) or dynamic programming. Reinforcement learning algorithms have shown the capacity to handle large and continuous state and action spaces, making them suitable for more complex problems [2, 8].

3.2 Reinforcement Learning

Reinforcement learning (RL) is a class of methods where an agent (or a decision maker) learns to make decisions by interacting with an environment, and receiving feedback in the form of rewards or penalties [31]. The environment is described by a MDP with deterministic states or by a partially observable MDP, where each state is represented by a distribution. The end goal of reinforcement learning is to maximize the cumulative discounted rewards obtained over a sequence of actions taken by the agent. In RL, the agent updates its knowledge of the environment based on the received rewards by relying on either an on-policy approach (e.g., SARSA and PPO) or an off-policy approach (e.g., Q-learning). On-policy methods refer to using the same policy to explore the environment and update the action-value function Q, while off-policy methods rely on two separate schemes for exploring the environment and updating the action-value function Q [31]. The action-value function Q provides estimates for the cumulative discounted rewards associated with each action at a given state. The estimation of Q-values can be done using the temporal difference (TD) [29]. The update rule in TD has different variations depending on the type of algorithm [27]. The Q-learning algorithm performs the updates on the Q-values using the observed reward and the estimated maximum Q-value of the next state-action pairs such that,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r(s_t, a_t) + \gamma \max_{a_{t+1} \in \mathcal{A}} (Q(s_{t+1}, a_{t+1})) - Q(s_t, a_t)], \tag{7}$$

Estimating the Q function can be done by either using tabular Q-learning or by relying on function approximation methods such as deep reinforcement learning.

Deep reinforcement learning (DRL) is a class of RL algorithms that relies on artificial neural networks (ANN) to approximate the Q-value function for each pair of state and action. One of the common DRL algorithms in the context of discrete action spaces is deep Q-network (DQN) [31]. The parameters $\boldsymbol{\theta}$ of the ANN model in DQN are updated recursively using the temporal difference (TD) approach [36], and by minimizing the loss function [15] defined as,

$$\mathcal{L}_i(\boldsymbol{\theta}_i) = \mathbb{E} \left[r(s, a) + \gamma\, Q(s', {}_{a'}\, Q(s', a'; \boldsymbol{\theta}_i); \boldsymbol{\theta}^-) - Q(s, a; \boldsymbol{\theta}_i) \right]^2, \tag{8}$$

where $\mathcal{L}_i(\boldsymbol{\theta}_i)$ is the loss function utilized for updating the parameters $\boldsymbol{\theta}_i$, and $\boldsymbol{\theta}^-$ is the target model parameters. The objective of the loss function in this context is to minimize the difference between the estimated Q-value at the current state s and the Q-value at the next state s', while also adding the observed immediate reward $r(\cdot)$. The target model in this loss function provides stability during the training process of the DRL agent, where the parameters of the target model $\boldsymbol{\theta}^-$ are updated by using either a hard update or soft update [18]. Another variation of DRL methods is the Dueling DQN which decomposes the Q-value function into two components,

$$Q(s_t, a_t; \boldsymbol{\theta}_v, \boldsymbol{\theta}_a) = V(s_t; \boldsymbol{\theta}_v) + A_V(s_t, a_t; \boldsymbol{\theta}_a), \tag{9}$$

where $A_V(s, a)$ is the estimate for the advantage of taking action a in state s, the set of parameters $\boldsymbol{\theta}_v$ correspond to the value function while the parameters $\boldsymbol{\theta}_a$ are associated with the advantage function. By separating these two components, the Dueling DQN can learn to distinguish the effects of different actions and states, which can improve the learning efficiency [35].

4 Quantifying the Costs of Delaying Maintenance

This section describes the formulation for a framework to quantify the cost of maintenance delays on the long-term maintenance costs.

4.1 Deviating from the Optimal Policy

In the presence of operational and budgetary constraints, it is challenging for decision-makers to perfectly follow the optimal maintenance policy required for each bridge in a network of bridges. Hence, this work focuses on the consequence of maintenance delays at the bridge-level. Specifically, if the maintenance is delayed for a duration of time $\mathtt{T}$, the loss function from Equation (2) can be simplified into,

$$\mathcal{L}_c(\boldsymbol{a}_t) = \sum_{t=0}^{\mathtt{T}} \gamma^t \sum_{p=1}^{\mathtt{E}} r_p(s_t, \bar{a}_t) + \sum_{t=\mathtt{T}+1}^{\infty} \gamma^t \sum_{p=1}^{\mathtt{E}} r_p(s_t, a_t^*), \tag{10}$$

with the action $\bar{a}_t$ refers to *do nothing*. Since no maintenance are performed during $\mathtt{T}$, the term $\sum_{t=0}^{\infty} \gamma^t \epsilon_t$ can be omitted in Equation (10). In order to make this formulation useful for decision makers, the loss function in Equation (10) is modified into,

$$\mathcal{L}_c(\boldsymbol{a}_t) = \bar{\mathbf{R}}_{t:\mathtt{T}} \times \sum_{t=0}^{\infty} \gamma^t \left(\sum_{p=1}^{\mathtt{E}} r_p(s_t, a_t^*) \right), \tag{11}$$

where $\bar{\mathbf{R}}_{t:\mathtt{T}}$ represents the relative change in the total maintenance cost due to deviation from the optimal maintenance policy up to time $\mathtt{T}$, defined by,

$$\bar{\mathbf{R}}_{t:\mathtt{T}} = \frac{\sum_{t=0}^{\mathtt{T}} \gamma^t \sum_{p=1}^{\mathtt{E}} r_p(s_t, a_t) + \sum_{t=\mathtt{T}+1}^{\infty} \gamma^t \sum_{p=1}^{\mathtt{E}} r_p(s_t, a_t^*)}{\sum_{t=0}^{\infty} \gamma^t \sum_{p=1}^{\mathtt{E}} r_p(s_t, a_t^*)}. \tag{12}$$

The variable $\bar{R}_t$ is defined in the range $[1, \infty]$, where $\bar{R}_t = 1$ implies that *doing nothing* is the optimal policy at time t, while $\bar{R}_t > 1$ represents a deviation from the optimal policy at time t. Accordingly, $\bar{R}_t$ provides decision makers with uniform and comparable insights about the relative change in the total cost required to maintain a bridge over time.

The estimation of $\bar{\mathbf{R}}_{t:\mathtt{T}}$ can be done by relying on concepts from reinforcement learning and more specifically Q-Learning. The Q function provides an approximation for the expected total discounted return associated with each state-action combination. Therefore, associating the return with monetary costs, can provide an interpretable meaning to the estimated Q values. For instance, if the return in the context of maintenance planning is represented by the cost associated with each maintenance action, then the Q function can be interpreted as an approximation for the total discounted costs (or NPV: net present value) for each state-action pair. This insight provides the foundations towards deriving the expected total costs for deviating from the optimal policy π^*.

Consider Q^* as the function associated with the optimal policy π^* for the action space $\mathcal{A}$ and state space $\mathcal{S}$. Deviating from the optimal policy at time t by taking a specific action $\bar{a}$, where $\bar{a}$ can be equal or different from $a \sim \pi^*(s_t)$, would incur an immediate cost Δ described by,

$$\Delta_t(s_t, a_t, \bar{a}) = Q(s_t, \bar{a}) - Q^*(s_t, a_t). \tag{13}$$

Generalizing Equation (13) for $\mathtt{T}$ time steps, the expected value for the incurred cumulative total discounted costs for deviating from π^* is,

$$\boldsymbol{\Delta}_{t:\mathtt{T}}(s_t, a_t, \bar{a}) = \sum_t \gamma^t \left[Q(s_t, \bar{a}) - Q^*(s_t, a_t)\right], \ \forall t = 0, \dots, \mathtt{T}, \tag{14}$$

where $\boldsymbol{\Delta}_{t:\mathtt{T}}$ is a vector containing the total expected change in costs due to deviating from the optimal policy, from time t up to time $\mathtt{T}$. From Equation (14), the relative change in the total maintenance cost $\bar{\mathbf{R}}_{t:\mathtt{T}}$ due to deviation from the optimal maintenance policy π^* up to time $\mathtt{T}$ is,

$$\bar{\mathbf{R}}_{t:\mathtt{T}}(s_t, a_t, \bar{a}) = \frac{Q^*(s_t, a_t) + \boldsymbol{\Delta}_{t:\mathtt{T}}(s_t, \bar{a})}{Q^*(s_t, a_t)}. \tag{15}$$

In the context of maintenance planning, $\bar{\mathbf{R}}_t$ corresponds to the relative change in the total maintenance cost for maintaining one structural element e_p^k. Accordingly, to evaluate a bridge-level cost ratio $\bar{\mathbf{R}}^b_{t:\mathtt{T}}$,

$$\bar{\mathbf{R}}^b_{t:\mathtt{T}}(s_t, a_t, \bar{a}, \boldsymbol{\omega}) = \frac{\sum_p^{\mathtt{P}} \omega_p \left[Q_p^*(s_t, a_t) + \boldsymbol{\Delta}_{t:\mathtt{T},p}(s_t, \bar{a})\right]}{\sum_p^{\mathtt{P}} [\omega_p \times Q_p^*(s_t, a_t)]}, \tag{16}$$

where $\mathtt{P}$ is the total number of elements within bridge $\mathcal{B}$. Equations (15) and (16) are employed for evaluating increments in total costs due to delaying maintenance actions on the bridge by assigning $\bar{a} = a_0 : \{\text{do nothing}\}$. The evaluation of $\bar{\mathbf{R}}_t$ and $\bar{\mathbf{R}}^b_t$, requires estimating the deterioration states s_t for each time t, as well as approximating the Q^* function, both of which are detailed in the following sections.

4.2 Estimating the Deterioration States

The estimation of the element-level deterioration states is done based on visual inspection data and by using state-space models (SSM) deterioration framework [10, 13]. The SSM framework describes the deterioration process over time using a transition model and an observation model. The transition model relies on a kinematic model [4], which describes the deterioration condition $x^k_{t,p}$, speed $\dot{x}^k_{t,p}$, and acceleration $\ddot{x}^k_{t,p}$ using,

$$\overbrace{\boldsymbol{x}^k_{t,p} = \mathbf{A}\boldsymbol{x}^k_{t-1,p} + \boldsymbol{w}_t}^{\text{transition model}}, \ \underbrace{\boldsymbol{w}_t : \boldsymbol{W} \sim \mathcal{N}(\boldsymbol{w}; \mathbf{0}, \mathbf{Q}_t)}_{\text{process errors}}. \tag{17}$$

The state vector at time t is represented by $\boldsymbol{x}^k_{t,p}$ where $\boldsymbol{x}_t : \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{x}, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ for the p-th element in the k-th structural category, $\mathbf{A}$ is the transition matrix, and $\boldsymbol{w}_t$ describes the process errors over time. The observation model on the other hand is described by,

$$\overbrace{y^k_{t,p} = \mathbf{C}\boldsymbol{x}^k_{t,p} + v_{t,i}}^{\text{observation model}}, \ \underbrace{v_{t,i} : V \sim \mathcal{N}(\boldsymbol{v}; \mu_V(I_i), \sigma^2_V(I_i))}_{\text{observation errors}}, \tag{18}$$

where $y^k_{t,p}$ is the observation on the deterioration condition of the structural element, $\mathbf{C}$ is the observation matrix, and $v_{t,i}$ is the observation error associated with each inspector $I_i \in \mathcal{I}$ responsible for the inspection task [10]. The transition of the deterioration states over time is performed using the Kalman filter (KF) and the RTS Kalman smoother (KS) [17, 26], and the monotonicity of the deterioration process is asserted by constraining the deterioration speed estimates, such that: $\dot{\mu}_{t|t} + 2\sigma^{\dot{x}}_{t|t} \leq 0$ [10]. The aforementioned criterion is examined at each time step t, and is enacted using the PDF truncation method [30]. Furthermore, it is possible to integrate the structural attributes (e.g., material) in the deterioration analyses by combining the SSM model with kernel regression (KR) [11].

Figure 1 shows an end-to-end diagram that describes the steps in the SSM-KR framework. The SSM-KR takes as an input the inspection data $\tilde{y}^k_{t,p}$, in addition to the material, structure age, latitude and health condition. The inspections $\tilde{y}^k_{t,p}$ are transformed to the unbounded space using the transformation function $o(\cdot)$, while the KR framework produces an initial estimate for the deterioration speed $\dot{x}^k_{0,p}$ based on the structural attributes. By using $y^k_{t,p}$ and $\dot{x}^k_{0,p}$, the SSM model produces the deterioration states $\boldsymbol{x}^k_{t,p}$ at each time t, which are transformed to the bounded space using inverse transformation function $o^{-1}(\cdot)$ to obtain the state $\tilde{\boldsymbol{x}}^k_{t,p}$ for interpretability [11]. Further details about the SSM-KR deterioration model formulation are available in the work of Hamida and Goulet [11, 13]

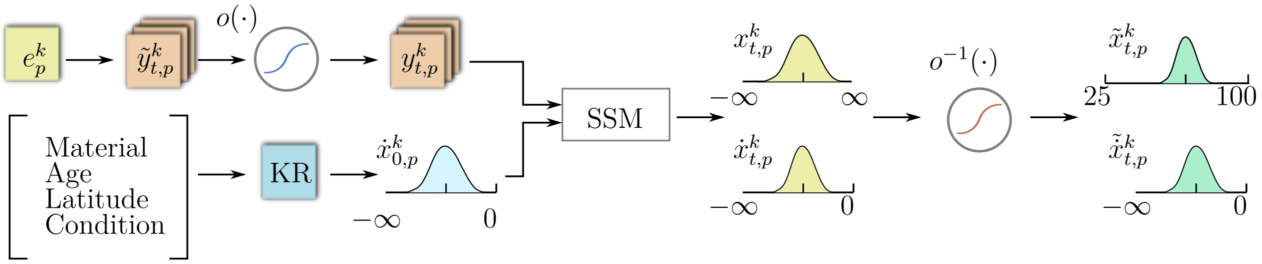

Figure 1: End-to-end diagram of the SSM-KR framework which takes as an input the inspection data $\tilde{y}^k_{t,p}$ and the structural attributes represented by the material, structure age, latitude and health condition. The inspection data $\tilde{y}^k_{t,p}$ are transformed from the bounded space to the unbounded space represented by $y^k_{t,p}$, while the structural attributes are processed in the KR to produce the initial deterioration speed estimate $\dot{x}^k_{t,p}$. The SSM framework takes the input and produces the deterioration state estimates in the unbounded space, which are thereafter back-transformed to the bounded space.

4.3 RL Environment and Agent Training

A RL environment represents a simulated setting in which an agent operates to learn and improve its decision-making. The InfraPlanner RL environment is composed of a pre-trained generative deterioration model based on the SSM-KR framework, a state aggregation model, and cost functions corresponding to the different types of structural elements and maintenance actions [14]. The role of the deterioration model is to generate realizations for the deterioration process by starting from a random initial health condition while taking into account the structural attributes of each structural element (e.g., material). The trajectory of a deterioration process realization remains susceptible to changes at any given time t via actions $a \in \mathcal{A}$ at the element level. An example of action is performing repairs on a beam structural element. Upon executing action a, the InfraPlanner environment incurs a cost defined in $[0, -\infty]$, contingent upon the type of action, the specific structural element involved, and the health condition of the element. Further details about the cost function are provided in Appendix A. The interactions between the RL agent and the environment are repeated recursively until the stopping criteria is met for the maximum number of interactions.

4.4 Quantifying the relative Change in Maintenance Costs

Figure 2 illustrates the full framework for quantifying the relative change in the expected costs $\bar{\mathbf{R}}^b_t$ associated with maintenance delays on a bridge $\mathcal{B}$. In this context, estimating $\bar{\mathbf{R}}^b_t$ at any time t requires taking into account the inspection data $\tilde{y}^k_{t,p}$ and the element quantity ω^k_p for each structural element e^k_p. The element quantity refers to either the unit size, such as the volume or the number of units within one element. The inspection data $\tilde{y}^k_{t,p}$ is analyzed using the SSM-KR deterioration model, which provides the deterioration state estimate $\tilde{\boldsymbol{x}}^k_{t,p}$, encompassing information about the deterioration condition $\tilde{x}^k_{t,p}$ and deterioration speed $\dot{\tilde{x}}^k_{t,p}$. From $\tilde{\boldsymbol{x}}^k_{t,p}$ a sample per element e^k_p is taken to represent a possible deterioration trajectory for each structural element, where $\boldsymbol{s}_t = [\tilde{x}^k_{t,p}, \dot{\tilde{x}}^k_{t,p}]$. Based on the state $\boldsymbol{s}_t$, it is possible to obtain the optimal action a^*_t, and the cost of deviating from the optimal policy Δ_t for the element e^k_p.

By repeating the steps for each element in the bridge and factoring the element quantity ω^k_p it is possible to estimate a single realization for the relative cost $\bar{R}^b_t$. In order to obtain multiple realizations for the relative cost, the aforementioned process is repeated $\mathtt{N}$ times, such that $\mathtt{N}$ possible deterioration trajectories are considered for each element in the bridge.

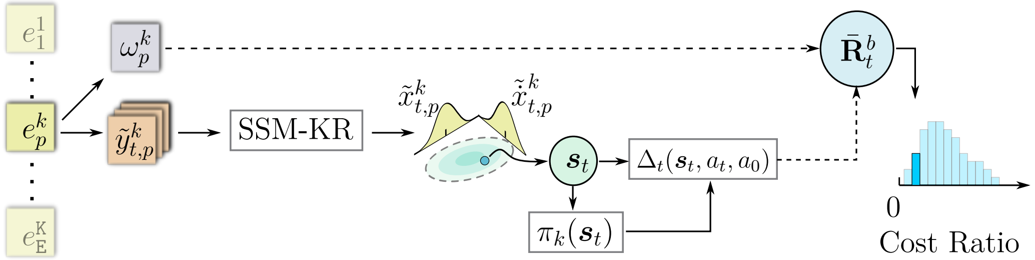

Figure 2: Estimating a single realization for the cost ratio $\bar{\mathbf{R}}_t^b$ based on the elements e_p^k from bridge $\mathcal{B}$. The estimation procedure starts with the inspection data $\tilde{y}_{t,p}^k$ which is analyzed using the SSM-KR deterioration model to provide the deterioration state $\tilde{\boldsymbol{x}}_{t,p}^k$ at time t. A sample from $\tilde{\boldsymbol{x}}_{t,p}^k$ is seeded into the state of the element $\boldsymbol{s}_t$ which alongside the normalized quantity ω_p^k enable estimating the cost of deviation Δ_t from the optimal policy π_k and the cost ratio $\bar{\mathbf{R}}_t^b$.

5 Example of Application

This section presents an application using the proposed relative cost evaluation method for quantifying the costs associated with delaying maintenance actions at the structural-element and the bridge level. The deterioration and intervention analyses presented in this section are based on a pre-trained SSM-KR deterioration model and an open-source RL environment *InfraPlanner* [14], which are calibrated based on the inspection data from the network of bridges in the Quebec province, Canada [11].

5.1 Bridge Description

The analyses in this case study are based on a concrete bridge $\mathcal{B}$ within the province of Quebec, Canada. The deterioration state of bridge $\mathcal{B}$ is monitored through visual inspections, which are performed every 3 years. The scope of the analyses includes only $\mathtt{K} = 6$ structural categories in bridge $\mathcal{B}$, which are, the beams $\mathcal{C}_1$, front-walls $\mathcal{C}_2$, slabs $\mathcal{C}_3$, wing-walls $\mathcal{C}_4$, guardrails $\mathcal{C}_5$, and pavement $\mathcal{C}_6$. The structural characteristics for each category are provided in Table 1.

Table 1: Characteristics of bridge $\mathcal{B}$ components, which includes the total number of elements in each structural category, the material, and the normalized element's quantity $\left(\text{i.e., } \left[\frac{\text{Quantity}}{\text{Min. Quantity}}\right]_{\times\text{Num. elements}}\right)$.

Category	# Elements	Material	Normalized Quantity
$\mathcal{C}_1$ Beams	15	Regular concrete	$[1]_{\times 10}, [2.57]_{\times 5}$
$\mathcal{C}_2$ Front Wall	2	Regular concrete	$[1]_{\times 1}, [1.39]_{\times 1}$
$\mathcal{C}_3$ Slabs	3	Regular concrete	$[1]_{\times 2}, [1.85]_{\times 1}$
$\mathcal{C}_4$ Wing Wall	4	Regular concrete	$[1]_{\times 1}, [1.2]_{\times 2}, [1.4]_{\times 1}$
$\mathcal{C}_5$ Guardrail	2	Wood/Steel	$[1]_{\times 1}, [1.25]_{\times 1}$
$\mathcal{C}_6$ Pavement	3	Asphalt	$[1]_{\times 2}, [1.84]_{\times 1}$

The above-mentioned structural categories can be classified based on their structural role within the bridge into two groups: 1) principal elements, and 2) secondary elements. Principal elements correspond to structural elements that carry or transfer structural loads (e.g., beams, front walls and slabs), while secondary elements correspond to elements associated with the serviceability of the bridge (e.g., wing walls, guardrails and pavement) [25]. For each structural group, there exists a predefined condition threshold where the health state is considered critical. If the deterioration state of any structural element goes below this threshold, immediate maintenance becomes necessary [25]. The condition threshold in this study is considered at $\tilde{x}_t = 50, \tilde{\dot{x}} = -1.8$ for the principal elements, and $\tilde{x}_t = 45, \tilde{\dot{x}} = -1.8$ for the secondary elements.

5.2 Learning the Maintenance Policies

The maintenance policies are learned by relying on the InfraPlanner RL environment, which emulates the deterioration condition and speed of different types of structural elements [14]. The RL environment takes as an input a maintenance action from the set $\mathcal{A}$, and returns a new state s_{t+1} at time $t+1$ in addition to the corresponding cost. The set of actions $\mathcal{A}$ is defined by $\mathcal{A} = \{a_0, a_1, a_2, a_3, a_4\}$, where a_0: do nothing, a_1: routine maintenance, a_2: preventive maintenance, a_3: repair, and a_4: replace. The cost and effects associated with each maintenance action are described in Appendix A. The configuration of the environment includes considering deterministic deterioration states, represented by $s_t = [x^k_{t,p}, \dot{x}^k_{t,p}]$, and infinite planning horizon with a discount factor $\gamma = 0.97$. The RL environment is reinitialized with a random health condition every $\mathtt{T} = 100$ years, a time span deemed sufficiently long to typically justify a replacement action in bridges.

A comparison for the performance of three RL agents is performed, namely, asynchronous DQN agent, asynchronous Dueling agent, and asynchronous Proximal Policy Optimization (PPO) agent. It is important to note that here, the asynchronous PPO is employed primarily for benchmarking purposes, since the PPO algorithm does not directly estimate the Q function [28]. As for the asynchronous training, it refers to evaluating the RL agent on multiple instances of the RL environment in parallel. There are mainly two advantages in using asynchronous training, 1) de-correlating the training samples which improves stability and 2) alleviate the requirement to store past experiences in a replay buffer [24]. Each RL agent is trained over a span of 3×10^6 steps. Detailed hyper-parameters and the experimental configuration for each agent can be found in Appendix B.

The asynchronous agents performance in learning the optimal maintenance policy is demonstrated on a slab structural element. Figure 3 shows the average performance based on 5 different seeds by the asynchronous DQN agent (dashed line), the asynchronous Dueling agent (dotted line) and the asynchronous PPO agent (dash-dot line). From Figure 3, the asynchronous agents show a similar overall performance, with the PPO agent having the best performance on average with $\mu_{\text{PPO}} = 5.7$, followed by the DQN agent achieving a slightly better total discounted cost on average with $\mu_{\text{DQN}} = 6.4$ than $\mu_{\text{Dueling}} = 6.7$ by the Dueling agent. Nonetheless, all agents have achieved a stable and relatively similar performance before the end of training threshold at 3×10^6 steps.

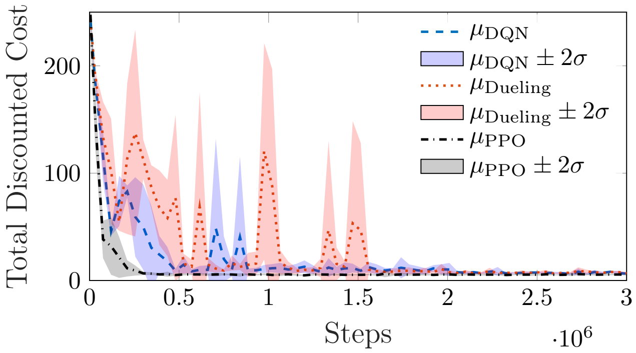

Figure 3: The training process of the asynchronous agents: DQN (dashed line), Dueling (dotted line) and PPO (dash-dot line) represented by the average performance in learning the optimal policy π^*_k for the slab structural element based using 5 different seeds.

Considering the near-optimal policy from the asynchronous DQN agent, it is possible to verify the accuracy of the agent in quantifying the total discounted costs by estimating it over a period of $\mathtt{T} = 100$ years, while starting from a fixed health state and following the policy π^*. For example, consider the RL environment to start in state $\boldsymbol{s}_0 = [\tilde{x}^k_{0,p}, \tilde{\dot{x}}^k_{0,p}]$, where $\tilde{x}^k_{0,p} = 55$ and $\tilde{\dot{x}}^k_{0,p}$ is obtained from the KR model as shown in Figure 1. The total discounted costs for being in state $\boldsymbol{s}_0$ can be estimated using,

$$\mu_{\text{Costs}} = \frac{1}{\mathtt{N}_s} \sum_{1}^{\mathtt{N}_s} \sum_{t=0}^{\mathtt{T}} \gamma^t r(s_t, \pi^*(s_t)), \tag{19}$$

where $\mathtt{N}_s$ is the total number of deterioration trajectories that started from state $\boldsymbol{s}_0$, and $r(\cdot)$ is the rewards function. Figure 4 shows a comparison between the optimal action-value function Q^* and the total discounted costs starting from different health condition states with $\mathtt{N}_s = 500$. From Figure 4, the optimal action-value function Q^* is within the confidence interval region of the total discounted costs estimate represented by $\mu_{\text{Costs}} \pm \sigma$, which implies that the Q^* function can be utilized to accurately estimate the total discounted costs while adhering the optimal policy π^*.

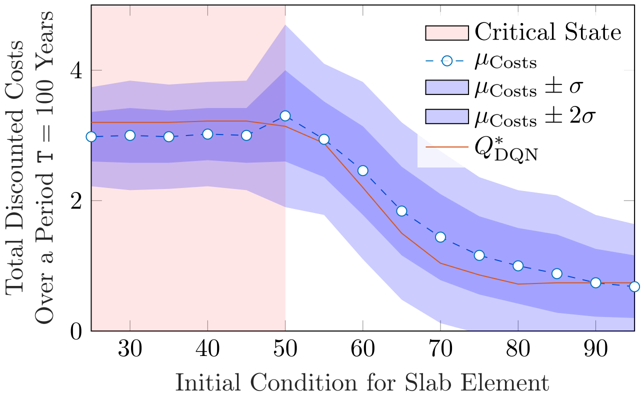

Figure 4: Comparison between the optimal action-value function Q^* and the total discounted costs estimated based on $\mathtt{N}_s = 500$ trajectories, over a period of $\mathtt{T} = 100$ years, for each health condition of the slab structural element.

5.3 Element Level Costs of Maintenance Delays

Following the estimation of the optimal maintenance policy, it becomes possible to obtain the Q value estimates associated with each maintenance action. Accordingly, the estimation of the relative cost $\bar{\mathbf{R}}_{t:\mathtt{T}}(s_t, a_t, \bar{a})$ using Equation (15) is feasible conditional to knowing the state s_t which is composed of the structural element condition $x^k_{t,p}$ and speed $\dot{x}^k_{t,p}$ at each time t. Obtaining estimates for the deterioration condition and speed can be done using the SSM-KR deterioration model, which relies on the visual inspection data and the structural attributes to provide an estimate for the deterioration state at each time step t. Figure 5 shows the deterioration analyses on a slab structural element e^3_1.

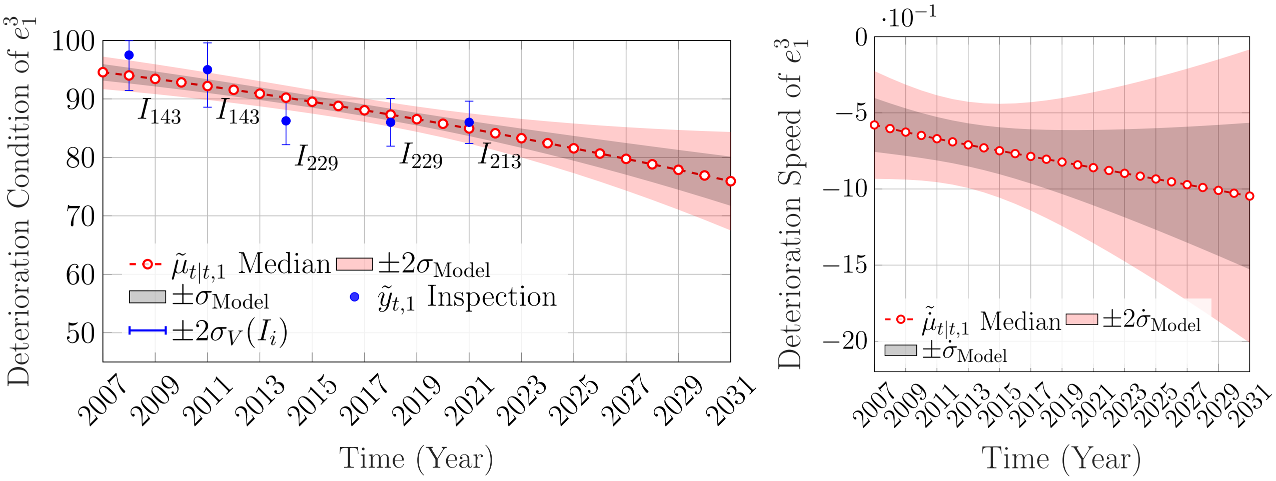

Figure 5: Deterioration state analysis for the condition and the speed based on the observations $\tilde{\boldsymbol{y}}^3_{t,1} \in [25, 100]$ of the structural element e^3_1 with the error bars representing the inspectors' uncertainty estimates.

In Figure 5, given the distribution of the deterioration state at year $t = 2021$, it possible to generate a sample to represent the state $\boldsymbol{s}_t$ as demonstrated in Figure 2. The sample represents the starting point of a single realization for the deterioration's trajectory. In this experiment, a total of $\mathtt{N} = 1000$ deterioration trajectories are generated and evaluated in Equation (15). Figure 6, shows the quantiles and median for the relative cost $\bar{\mathbf{R}}_{t:\mathtt{T}}(s_t, a_t, a_0)$ of element e^3_1. From Figure 6, the total discounted cost required for maintaining element e_1^3 starts to increase after the year 2027 due to delaying maintenance actions, and reaches a factor 1.5× the total discounted cost at the year 2031. Such information is useful when considering maintenance delays on this bridge.

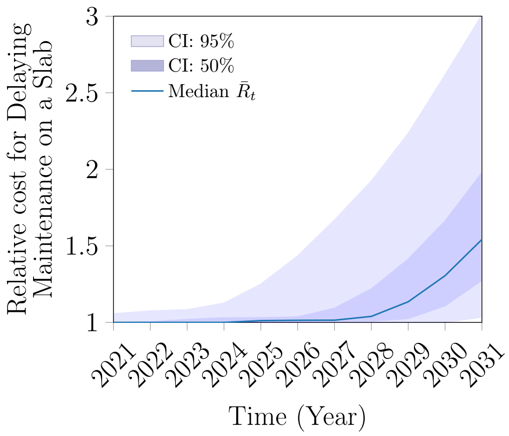

Figure 6: Relative cost $\bar{R}_t$ estimated based on the optimal policy π_3^* and a $\mathtt{N} = 1000$ realization of deterioration trajectories for a slab structural element e_1^3.

Another element-level example covers a scenario where a replacement action a_4 is considered as an optimal action by the maintenance policy π^*. The example is for a wing wall structural element e_1^4 in bridge $\mathcal{B}$. Figure 7 shows the deterioration condition and speed estimated based on visual inspection data over time. From Figure 7, there are noticeable drops in the condition in years 2011 and 2014, which coincided with a high deterioration speed over time.

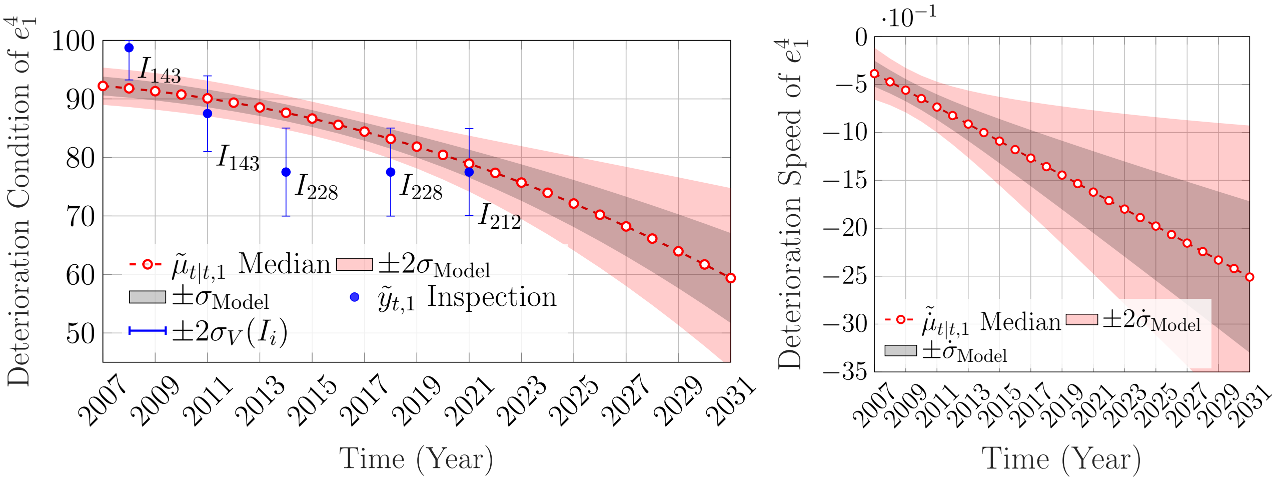

Figure 7: Deterioration state analysis for the condition and the speed based on the observations $\tilde{\boldsymbol{y}}_{t,1}^4 \in [25, 100]$ of the structural element e_1^4 with the error bars representing the inspectors' uncertainty estimates.

Based on the optimal policy for maintaining wing wall elements π_4^* shown in Figure 10, it is likely that a replacement action a_4 is required due to the high deterioration speed. Figure 8 illustrates the cost ratio and the probability of replacement represented by $\Pr(a^* = a_4)$ over time and based on $\mathtt{N} = 1000$ realization of deterioration trajectories for e_1^4.

In Figure 8, the probability $\Pr(a^* = a_4)$ is quantified by computing the ratio between the number of realizations where replacement is the optimal action $a^* = a_4$ at a given year t, and the total number of realizations $\mathtt{N} = 1000$. Accordingly, the probability of replacement increases over time to reach $\Pr(a^* = a_4) \approx 40\%$ at year 2031. It should be noted that the cost of a replacement action a_4 is added only once to the total discounted change in costs $\boldsymbol{\Delta}_{t:\mathtt{T}}$ in Equations 14 and 15. For example, if replacement is the optimal action at year t, then the total discounted change in cost at year $t+1$ is

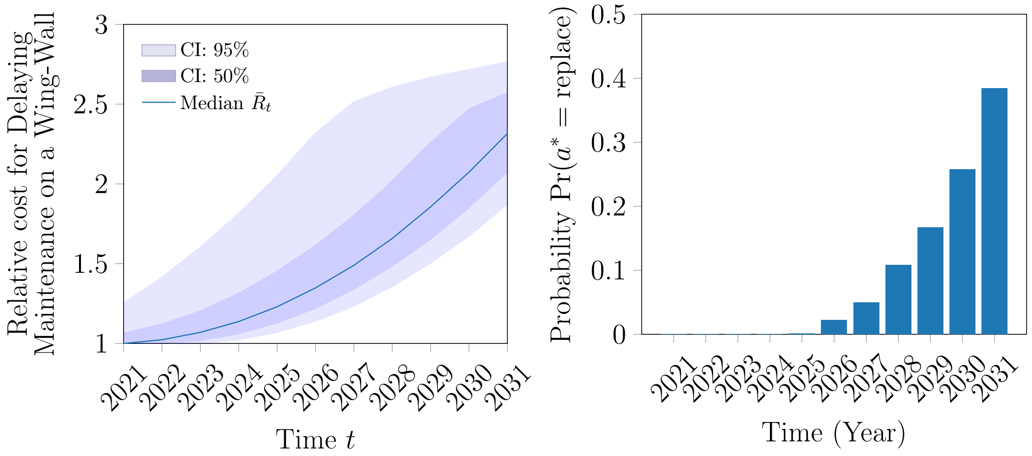

Figure 8: Relative cost $\bar{R}_t$ estimated based on the optimal policy π_4^* and a $\mathtt{N} = 1000$ realization of deterioration trajectories for a wing wall structural element e_1^4 (left), and the probability of replacement over time for the element e_1^4 (right).

$\Delta_{t+1} = 0$. This is done to avoid augmenting the cost ratio $\bar{R}_t$ with costs that are beyond the actual maximum cost, which is the cost of replacement in this context.

5.4 Bridge Level Costs of Maintenance Delays

As described in Section 5.1, the bridge $\mathcal{B}$ is composed of $\mathtt{K} = 6$ structural categories, with each structural category containing different number of elements. In order to provide an insight into the overall deterioration condition and speed of bridge $\mathcal{B}$, Figure 9 shows the deterioration states which are obtained by aggregating the element-level deterioration states using a Gaussian mixture reduction approach [13].

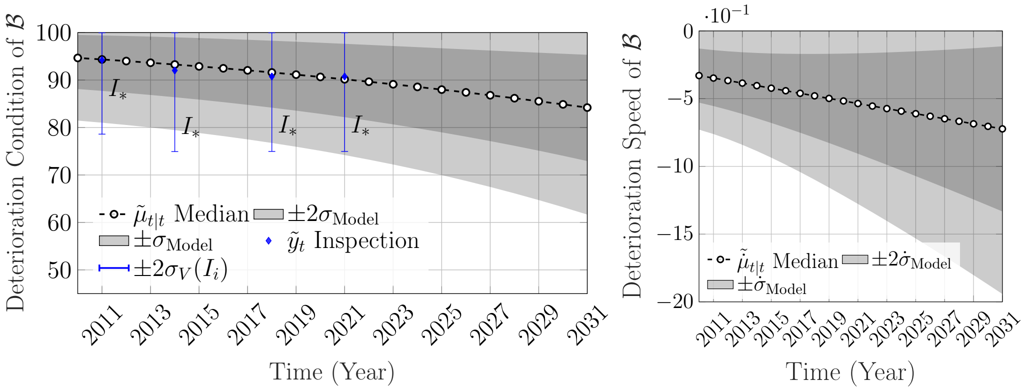

Figure 9: Deterioration state estimates for the condition and speed of bridge $\mathcal{B}$, based on the aggregation of element-level deterioration states of the structural categories $\mathcal{C}_{1:\mathtt{K}}$, with the aggregated inspections $\tilde{y}_t \in [25, 100]$ represented by the blue diamond, and their corresponding uncertainty estimates represented by the blue error bars.

For a bridge-scale application of the proposed relative cost approach, it is required to obtain an optimal maintenance policy for each structural category in the bridge $\mathcal{B}$. Figure 10 shows the optimal policy maps for each structural category in the bridge $\mathcal{B}$, as obtained by the asynchronous DQN agents. Each policy map represents a mapping between the deterioration state and maintenance actions. In this assessment, the policy maps are learned based on two condition thresholds for the elements. A risk-averse threshold (or high threshold) defined by $(\tilde{x}_t = 60, \tilde{\dot{x}} = -1.5)$ for the principal elements, and $(\tilde{x}_t = 55, \tilde{\dot{x}} = -1.5)$ for the secondary elements, and a risk-neutral threshold (or low threshold) defined by $(\tilde{x}_t = 50, \tilde{\dot{x}} = -1.8)$ for the principal elements, and $(\tilde{x}_t = 45, \tilde{\dot{x}} = -1.8)$ for the secondary elements. It should be noted that the risk-neutral thresholds align with the criteria defining critical health conditions for elements, as specified in the inspection manual [25].

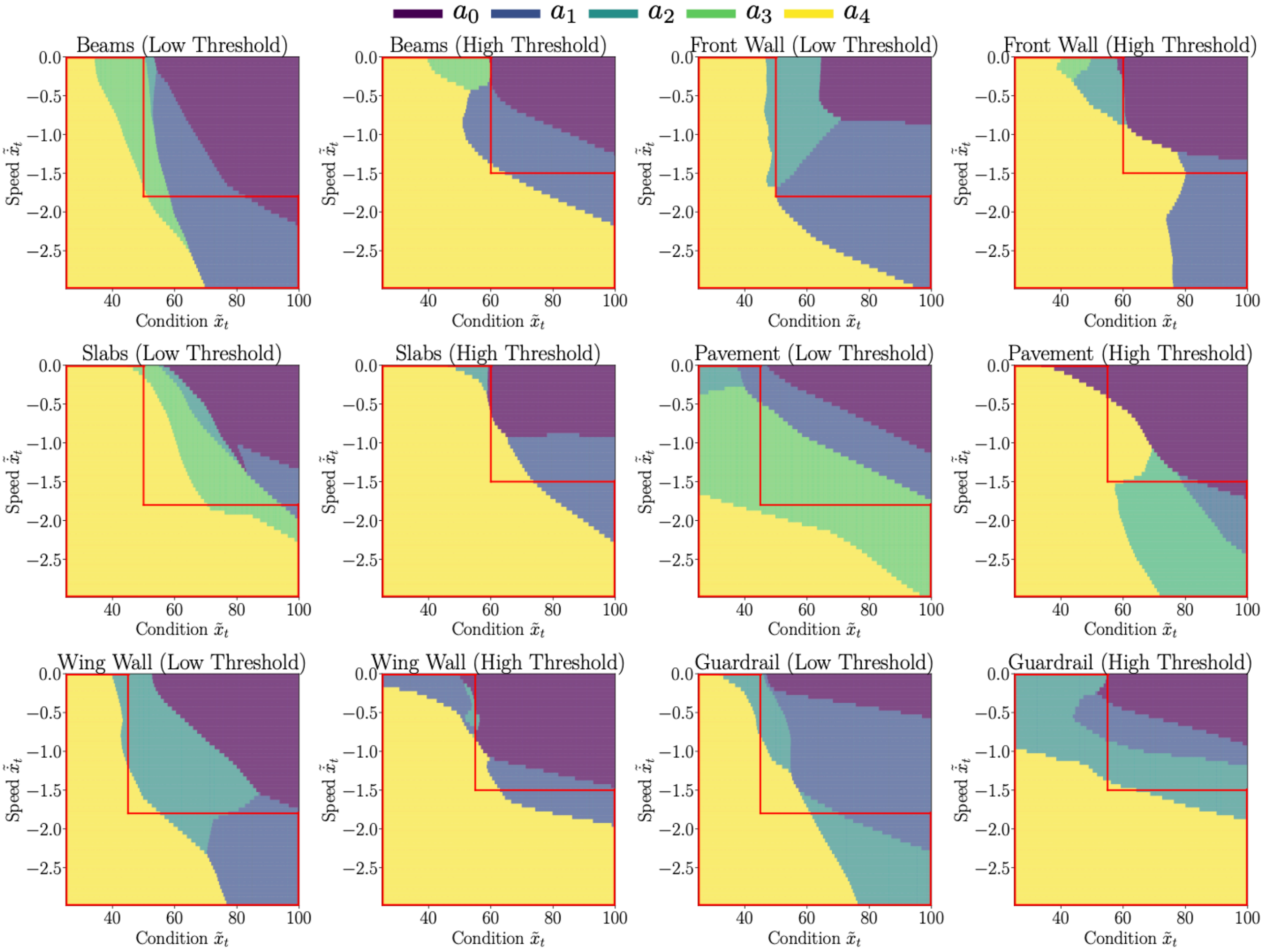

Figure 10: Maintenance policy maps obtained by asynchronous DQN agents for each structural category in bridge $\mathcal{B}$, where each graph represents a mapping between states and maintenance actions. The area enclosed within the red frame represents the predefined minimum condition thresholds for the condition $\tilde{x}^k_{t,p}$ and speed $\tilde{\dot{x}}^k_{t,p}$.

By relying on the optimal maintenance policies in Figure 10, and the framework described in Section 4.4, it becomes possible to estimate the overall bridge-level relative cost $\bar{R}^b_t$, which is shown in Figure 11. Based on Figure 11, delaying maintenance actions on the bridge $\mathcal{B}$ can result in increasing the total discounted costs of maintenance by a 1.1× factor, given the optimal maintenance policies obtained based on the low minimum condition thresholds. On the other hand, policy maps with higher condition thresholds have yielded a similar relative cost $\bar{R}^b_t$, with a slightly higher increase in the total discounted maintenance costs by a 1.15× factor due to maintenance delays. Despite the differences in the maintenance policies between the two scenarios (low threshold vs. high threshold), the relative cost $\bar{R}^b_t$ for each case are fairly similar. This is attributed to the fact that the framework is examining the relative costs rather than the total costs over time.

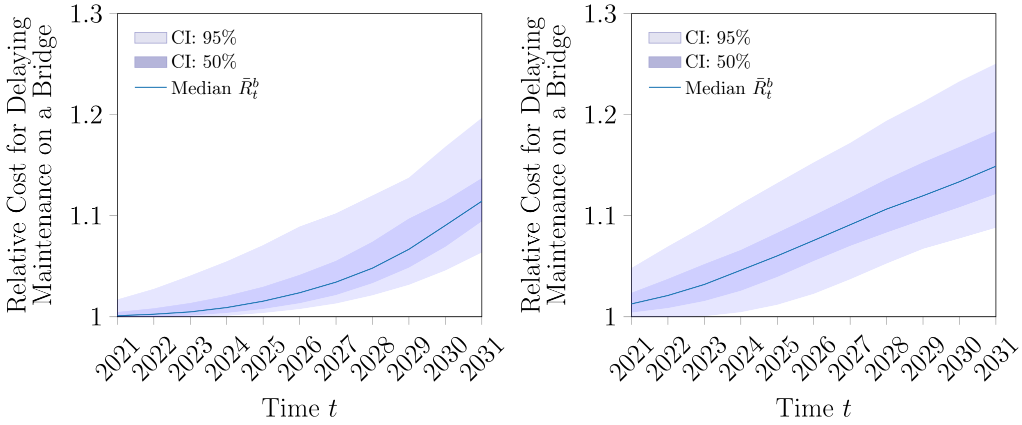

Figure 11: Relative cost $\bar{\mathbf{R}}_t^b$ estimated based on the optimal policy maps and a $\mathtt{N} = 1000$ realization of deterioration trajectories for each structural element in bridge $\mathcal{B}$. The graph on the left corresponds to the policy maps associated with the low condition thresholds, while the graph on the right corresponds to policy maps with high condition thresholds.

6 Conclusion

Bridge maintenance planning demands careful consideration of various factors which can be classified into two main groups: 1) structure-level factors such as the structural integrity and safety of the bridge, and 2) network-level factors such as the bridge's influence on the connectivity in the transportation network. This study concentrates on structure-level factors, with a specific focus on examining the influence of the deterioration process on the maintenance costs and decision-making. The proposed approach is based on a deep reinforcement learning framework, and aims at quantifying the cost ratio between the expected discounted costs of taking no action and the expected discounted costs of taking the optimal maintenance action. The aforementioned cost ratio is formulated to evaluate the incurred costs at the element-level and the bridge-level. Applications using the proposed approach is demonstrated using inspections data from a bridge within the province of Quebec, Canada. The analyses involved identifying an optimal maintenance policy for each structural category in the bridge using asynchronous deep Q network (DQN). The capacity of the asynchronous DQN agent to estimate the total discounted costs is verified through a comparison with a sampling-based approach. The verification results have demonstrated that the deep RL agent estimates of the total discounted costs are within the confidence interval of the sampling approach. The cost ratio estimates are shown at an element-level for a slab element and a wing wall element, where the probability of element replacement is quantified. This is followed by bridge-level estimates for the relative costs using the proposed approach. The results of the analyses have demonstrated the capacity of the relative cost metric in translating information about the gradual deterioration of elements into information about changes in maintenance costs. The results also highlight that the Q function values for actions other than the optimal action can provide valuable information, which in this case corresponds to information about the costs associated with maintenance delays. The proposed framework is shown to be scalable as the search for an optimal maintenance policy occurs at the element level, where the state space and action space are relatively small. Despite the aforementioned advantages, the relative cost metric is by design dependent on the optimal maintenance policy and the accuracy of the deterioration model estimates. Accordingly, verifying the performance of the RL agent and the deterioration model are prerequisites to ensure the reliability of the cost ratio metric. The proposed approach may not provide an exact scheduling solution, however, it offers valuable insights to decision-makers about the tradeoffs associated with maintenance delays. Future work in this context includes the use of distributional RL and partially observable POMDP environment, which would alleviate the need for sampling deterioration trajectories.

Acknowledgements

This project is funded by the Transportation Ministry of Quebec Province (MTQ), Canada. The authors would like to acknowledge the support of Simon Pedneault for facilitating the access to information related to this project.

A Effects and Costs of Maintenance Actions

The effect of maintenance actions are dependent on the type of structural category and are reproduced from the work of Hamida and Goulet [14]. The deterministic maintenance effects defined in Table 2 are derived from estimates based on visual inspection data from the network of bridges in the province of Quebec [12].

Table 2: Table of the true effects associated with maintenance actions on each structural category [14].

	Structural Category					
	Beams	Front Wall	Slabs	Guardrail	Wing Wall	Pavement
a_0	0	0	0	0	0	0
a_1	0.5	0.1	1	0.25	0.25	8
a_2	7.5	19	12	9	8	20
a_3	18.75	20.5	20	14	17	28
a_4	75	75	75	75	75	75

The cost of maintenance actions are defined as a function of the deterioration state [14], and are shown in Figure 12. The relation between the cost and the condition is described by $x_c(\tilde{x}^k_{t,p}, a) = \beta_1(a)\frac{1}{\tilde{x}^k_{t,p}} + \beta_2(a)$ [14]. Accordingly, the total cost r at any time t is,

$$r(\boldsymbol{s}_t, a_t) = x_c(\tilde{x}^k_{t,p}, a_t) + r^p,$$

where r^p is a fixed cost representing a penalty applied on the agent if the critical condition threshold is reached. Further details about the cost function are available in the work of Hamida and Goulet [14].

B Asynchronous DRL Hyper-parameters

The asynchronous RL agents are trained on InfraPlanner RL environment using a discount factor 0.97. A total of $\mathtt{n} = 50$ copies of randomly seeded environments are evaluated by the asynchronous RL agent. The setup of both the dueling and DQN agents involves the use of ϵ-greedy for exploration. The ϵ value is decayed linearly over the first 200 episodes to a minimum value 0.01. The target model updates are performed every 1000 steps in the environment. All neural networks have the same architecture consisting in 2 layers of 128 hidden units and $relu(\cdot)$ activation functions. The learning rate starts at 10^{-3} and is decayed to 10^{-5} after 800 episodes. On the other hand, the PPO agent has the same neural network architecture with $tanh(\cdot)$ activation functions, and a constant learning rate at 10^{-4}.

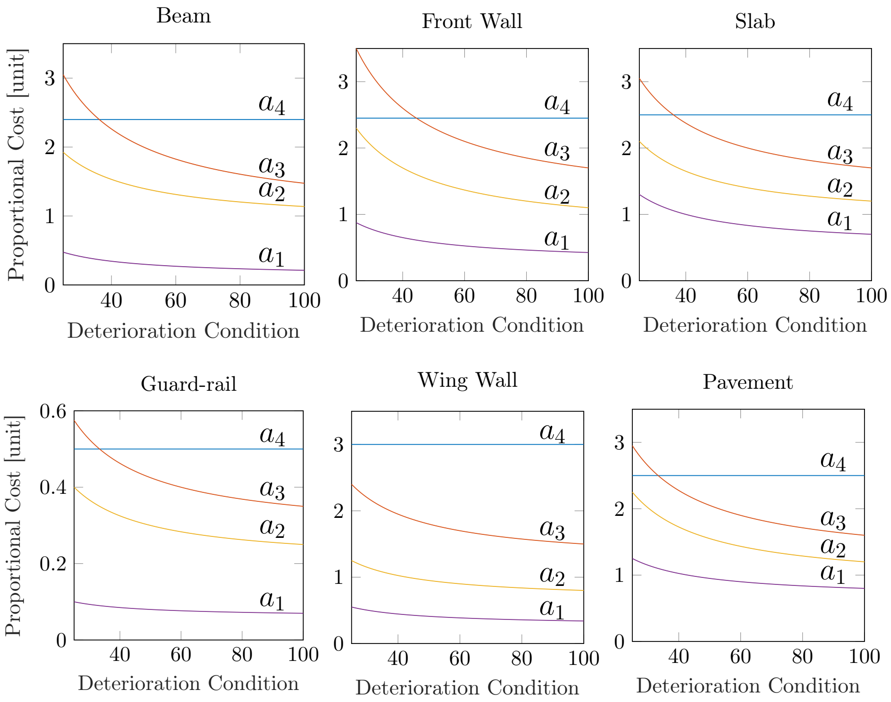

Figure 12: Proportional costs of each maintenance action as a function of the deterioration condition [14].

References

[1] Zaharah Allah Bukhsh, Irina Stipanovic, and Andre G Doree. Multi-year maintenance planning framework using multi-attribute utility theory and genetic algorithms. *European transport research review*, 12(1):1–13

[2] Charalampos P Andriotis and Konstantinos G Papakonstantinou. Managing engineering systems with large state and action spaces through deep reinforcement learning. *Reliability Engineering & System Safety*, 191:106483, 11 2019. ISSN 09518320. doi: 10.1016/j.ress.2019.04.036. URL `https://linkinghub.elsevier.com/retrieve/pii/S0951832018313309`.

[3] Atorod Azizinamini, Edward H. Power, Glenn F. Myers, H. Celik Ozyildirim, Eric S. Kline, David W. Whitmore, and Dennis R. Mertz. Design guide for bridges for service life, chapter 11 life-cycle cost analysis. Technical report, The National Academies Press, Washington, DC, 2013.

[4] Yaakov Bar-Shalom, X Rong Li, and Thiagalingam Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software.* John Wiley & Sons, 2004.

[5] Taeyeon Chang, Gyueun Lee, and Seokho Chi. Development of an optimized condition estimation model for bridge components using data-driven approaches. *Journal of Performance of Constructed Facilities*, 37(3):04023013

[6] Min-Yuan Cheng, Yi-Cho Fang, Yung-Fang Chiu, Yu-Wei Wu, and Ting-Chang Lin. Design and maintenance information integration for concrete bridge assessment and disaster prevention. *Journal of Performance of Constructed Facilities*, 35(3):04021015

[7] Cristian Contreras-Nieto, Yongwei Shan, Phil Lewis, and Julie Ann Hartell. Bridge maintenance prioritization using analytic hierarchy process and fusion tables. *Automation in Construction*, 101: 99–110, 5 2019. ISSN 09265805. doi: 10.1016/j.autcon.2019.01.016.

[8] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468 0885–6125, 2021.

[9] Georgios M Hadjidemetriou, Manuel Herrera, and Ajith K Parlikad. Condition and criticality-based predictive maintenance prioritisation for networks of bridges. *Structure and Infrastructure Engineering*, 18(8):1207–1221

[10] Zachary Hamida and James-A Goulet. Modeling infrastructure degradation from visual inspections using network-scale state-space models. *Structural Control and Health Monitoring*, pages 1545–2255, 2020. doi: 10.1002/stc.2582.

[11] Zachary Hamida and James-A Goulet. Network-scale deterioration modelling based on visual inspections and structural attributes. *Structural Safety*, 88:102024, 2020. doi: 10.1016/j.strusafe.2020.102024.

[12] Zachary Hamida and James-A. Goulet. Quantifying the effects of interventions based on visual inspections of bridges network. *Structure and Infrastructure Engineering*, pages 1–12, 2021. doi: 10.1080/15732479.2021.1919149.

[13] Zachary Hamida and James-A. Goulet. A stochastic model for estimating the network-scale deterioration and effect of interventions on bridges. *Structural Control and Health Monitoring*, pages 1545–2255, 2021. doi: 10.1002/stc.2916.

[14] Zachary Hamida and James-A Goulet. Hierarchical reinforcement learning for transportation infrastructure maintenance planning. *Reliability Engineering & System Safety*, 2023.

[15] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence.*, volume 30, 2016.

[16] Hugh Hawk. Bridge life-cycle cost analysis guidance manual. Technical Report 483, National Cooperative Highway Research Program, 2003.

[17] Rudolf Emil Kalman. Contributions to the theory of optimal control. *Bol. Soc. Mat. Mexicana*, 5 (2):102–119, 1960.

[18] Taisuke Kobayashi and Wendyam Eric Lionel Ilboudo. T-soft update of target network for deep reinforcement learning. *Neural Networks*, 136:63–71

[19] Xiaoming Lei, Ye Xia, Lu Deng, and Limin Sun. A deep reinforcement learning framework for life-cycle maintenance planning of regional deteriorating bridges using inspection data. *Structural and Multidisciplinary Optimization*, 65, 5 2022. ISSN 16151488. doi: 10.1007/s00158-022-03210-3.

[20] Zhenglin Liang and Ajith Kumar Parlikad. Predictive group maintenance for multi-system multi-component networks. *Reliability Engineering & System Safety*, 195:106704

[21] Min Liu and Dan M. Frangopol. Bridge annual maintenance prioritization under uncertainty by multiobjective combinatorial optimization. *Computer-Aided Civil and Infrastructure Engineering*, 20:343–353, 9 2005. ISSN 10939687. doi: 10.1111/j.1467-8667.2005.00401.x.

[22] Zhongxiang Liu, Tong Guo, José Correia, and Libin Wang. Reliability-based maintenance strategy for gusset plate connections in steel bridges based on life-cost optimization. *Journal of Performance of Constructed Facilities*, 34(5):04020088

[23] Bruce L Miller. *Finite state continuous time Markov decision processes with an infinite planning horizon.* Rand Corporation, 1967.

[24] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[25] MTQ. *Manuel d'Inspection des Structures.* Ministère des Transports, de la Mobilité Durable et de l'Électrification des Transports, Jan 2014.

[26] Herbert E Rauch, CT Striebel, and F Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450 0001–1452, 1965.

[27] Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[29] Mohit Sewak and Mohit Sewak. Temporal difference learning, sarsa, and q-learning: Some popular value approximation based reinforcement learning approaches. *Deep Reinforcement Learning: Frontiers of Artificial Intelligence*, pages 51–63

[30] Dan Simon and Donald L Simon. Constrained kalman filtering via density function truncation for turbofan engine health estimation. *International Journal of Systems Science*, 41(2):159–171

[31] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

[32] Adam J. Trzcinski and Ross B. Corotis. Alternative valuation of highway user delay costs. *Civil Engineering and Environmental Systems*, 24:87–97, 6 2007. ISSN 10286608. doi: 10.1080/10286600601156632.

[33] Sergio Valenzuela, Hernan de Solminihac, and Tomas Echaveguren. Proposal of an integrated index for prioritization of bridge maintenance. *Journal of Bridge Engineering*, 15:337–343, 5 2010. ISSN 1084-0702. doi: 10.1061/(asce)be.1943-5592.0000068.

[34] Hao Nan Wang, Ning Liu, Yi yun Zhang, Da wei Feng, Feng Huang, Dong sheng Li, and Yi ming Zhang. Deep reinforcement learning: a survey. *Frontiers of Information Technology and Electronic Engineering*, 21:1726–1744, 12 2020. ISSN 20959230.

[35] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.

[36] Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards.* PhD thesis, University of Cambridge, King's College, Cambridge United Kingdom, 1989.

[37] Shiyin Wei, Yuequan Bao, and Hui Li. Optimal policy for structure maintenance: A deep reinforcement learning framework. *Structural Safety*, 83:101906

[38] David Y Yang and A M Asce. Deep reinforcement learning-enabled bridge management considering asset and network risks. *Journal of Infrastructure Systems*, 2022.

[39] Nailong Zhang and Wujun Si. Deep reinforcement learning for condition-based maintenance planning of multi-component systems under dependent competing risks. *Reliability Engineering and System Safety*, 203, 11 2020. ISSN 09518320. doi: 10.1016/j.ress.2020.107094.

[40] Weili Zhang and Naiyu Wang. Bridge network maintenance prioritization under budget constraint. *Structural Safety*, 67:96–104 URL `https://doi.org/10.1016/j.strusafe.2017.05.001`.