Modeling the Business Value of a Predictive Maintenance System using Monte Carlo Simulation

Graeme Garner[1], Paola Santanna[2], and Hossein Sadjadi[3]

[1,3]*General Motors Company, Canadian Technical Center, Markham, Ontario, L3R 4H8, Canada*

graeme.garner@gm.com

hossein.sadjadi@gm.com

[3]*General Motors Global Technical Center, Warren, Michigan, 48093, USA*

paola.santanna@gm.com

ABSTRACT

The automotive industry is undergoing a period of rapid advancement, as original equipment manufacturers race to develop the next generation of electric, autonomous, and connected vehicles. Many manufacturers are investing in prognostics technology, which has made advancements mainly in the aerospace industry over the past couple decades. For vehicle fleet managers who own and operate many vehicles, prognostics and early fault detection can enable predictive maintenance strategies, which can realize cost savings versus corrective or preventative strategies. However, developing the technology required for predictive maintenance can be an expensive undertaking, requiring many parts, months of data collection, and possibly years of engineering effort. It is critical to understand the expected return on investment for developing such a project.

In this paper, we present a framework to model the business value of a predictive maintenance system. The predictive maintenance system is described as the combination of a component being monitored, a network of sensors, a health monitoring algorithm, and a service policy defining the response to those actions. The framework incorporates models of component failure, health monitoring algorithm performance, a policy of actions, and costs associated with those actions. The framework is generic and may be applied to any component where degradation can be modelled by a probability distribution. Monte Carlo simulation is employed to estimate the distribution of repair costs for a particular maintenance strategy, which can then be used to assess the value of a predictive maintenance system.

1. INTRODUCTION

As automotive sensor and controller technology developed over the past handful of decades, the responsibility for fault detection has shifted from vehicle owners and technicians listening for abnormal sounds and taking hand measurements to on-board computerized systems that issue automatic warnings. All modern vehicles are equipped with an on-board diagnostics (OBD) port, which allows any driver, fleet manager, or technician to plug in a tool and analyze the health of dozens of vehicle components by looking at diagnostic trouble codes (DTCs). The available trouble codes and expected OBD functionalities are captured by international standards, such as J1979 (SAE, 2012) and J2012 (SAE, 2013). For the most part, the rise of OBD has been driven by government bodies, such as the California Air Resources Board (CARB) which mandates a minimum level of emissions-related diagnostics that must exist on a vehicle for it to be sold in their jurisdiction (Cal. Code of Regulations, 2021). Other diagnostics are required to maintain the safety of a vehicle, such as those related to wheel speed sensor, brakes, and active safety or autonomous features.

There is, however, incentive for automotive original equipment manufacturers (OEMs) to develop diagnostics beyond the government requirements and safety responsibilities. Early fault detection enables owners to fix problems before they cause costly compound effects. This benefits the owner of the vehicle as it can prevent a walk-home scenario, and it also benefits the OEM by reducing cost in cases where the repairs are covered under warranty. Fleet managers are particularly interested in these systems because they enable optimization of fleet maintenance and prevention of costly downtime (Fuchs, Safar, & Kok, 2016).

Typically, a component on a vehicle is covered by one of three service strategies: corrective, preventative, or predictive. Corrective maintenance involves only replacing a

part when a fault has already occurred. This is most common for components that are neither critical nor expected to wear. Preventative maintenance involves regular servicing or replacement to mitigate the risk of an in-service failure. Regular oil changes are an example of commonly practiced preventative maintenance for automobiles. Finally, predictive maintenance is any strategy that services components based on their predicted condition. This strategy is commonly referred to as "condition-based" maintenance, as it requires a condition monitoring system to assess component health.

Defining the maintenance strategy for a vehicle is a complex optimization problem as each strategy has pros and cons. Corrective maintenance typically has the lowest short-term costs since only failed parts are replaced. However, in-service failures can lead to costly downtime and compound effects that can drive greater costs in the long term. Preventative maintenance seeks to mitigate random, unexpected costs and downtime with a regular schedule of known costs. Over-scheduling maintenance, however, can lead to preventative maintenance costs exceeding corrective maintenance costs. Predictive maintenance presents an opportunity to reduce the overall cost of preventative maintenance while maintaining the benefit of reduced risk of downtime. However, implementing a predictive maintenance strategy requires a fault detection system, which is usually neither easy nor cheap to develop. Savings are not guaranteed, since placing trust in a fault detection system with poor performance can lead to a predictive strategy having greater cost than a preventative or corrective one.

There is evidence of predicative maintenance yielding positive results in certain applications. McKinsey has published evidence of a 20% reduction in downtime achieved by predictive maintenance, although they also highlighted the massive development effort required to achieve that benchmark (Decaix, Gentzel, Luse, Neise, & Thibert, 2021). Deloitte claims similar figures in their position paper (Deloitte Analytics Institute, 2017), citing an average reduction in costs by 25% and reduction in breakdowns by 70%. Both firms cite the clear fact that the value of applying preventative maintenance to a component is correlated with the maintenance costs and downtime caused by that component, as well as with the performance of the condition monitoring technology.

Maintenance cost modelling and optimization has been widely studied, and many academic works have been published on mathematical formulations and approaches to solving this problem. In Alrabghi and Tiwari's literature review (2013), a large body of simulation-based maintenance optimization methods are explored. Methods employing discrete-event simulation were found to be most common, supporting earlier findings of Jahangirian, Eldabi, Aisha, Stergioulas, and Young (2010). The problem of determining optimal service frequency for preventative maintenance has been deeply explored (Khandelwal, Sharma, & Ray, 1979).

In their review on maintenance optimization, Jonge and Scarf (2019) noted a recent trend in extending models to account for condition-based (predictive) strategies. They broadly classify research addressing single-component and multi-component systems, and those assuming perfect and imperfect maintenance results. Most relevant to the focus of this paper is the work of Xiang, Cassady, and Pohl (2012), who use a Markovian model to assess the cost savings of a predictive maintenance strategy that accounts for error in the condition monitoring algorithm. Their approach assumes a prognostic with gaussian error is used to drive maintenance decisions.

For OEMs seeking to optimize the maintenance costs for their vehicles, the problem comes in identifying which components are worth the effort to develop a fault detection system to enable predictive maintenance, and which are better served by a corrective or preventative strategy. As outlined in the literature review of Jonge and Scarf (2019), there are a plethora of models available to simulate and optimize the maintenance strategy for any given component. Careful thought and engineering judgement must be applied when selecting which model to apply to a given component, and not all components will warrant the same approach. The problem this paper aims to resolve is the lack of uniformity in the approach to valuating maintenance strategies. A flexible framework that can be applied to any component or system will benefit organizations that need to compare many options for focusing their engineering efforts.

In this paper, we describe such a flexible framework that enables simulation-based valuation of maintenance strategies. The components of the framework are described mathematically, and a simple example is shared to highlight the insights that this framework can enable.

2. VALUATION FRAMEWORK

2.1. Predictive Maintenance

As this paper defines a valuation framework for a predictive maintenance system (PMS), we begin by precisely defining that system and its components. The goal of a PMS is to monitor the health of a component on a vehicle, and perform maintenance as required. For the remainder of this paper, we will use the term "vehicle", but the ideas will apply to any machine or system with serviceable components.

As summarized in Figure 1, a PMS comprises of a network of sensors to monitor the vehicle and component, a health monitoring algorithm (HMA), and a service policy that defines actions to be taken given the outputs of the algorithm. The value of a PMS comes from the service actions it takes, so the system includes both the health monitoring technology itself plus the policy that defines the actions taken in response to that technology.

The goal of the HMA is to estimate the state-of-health (SOH) of the component. There are many possible realizations of an HMA – a simple diagnostic HMA may estimate the SOH as a Boolean (healthy or faulty), or more advanced prognostics HMAs may estimate both the current SOH on a continuum and the estimated remaining useful life (RUL).

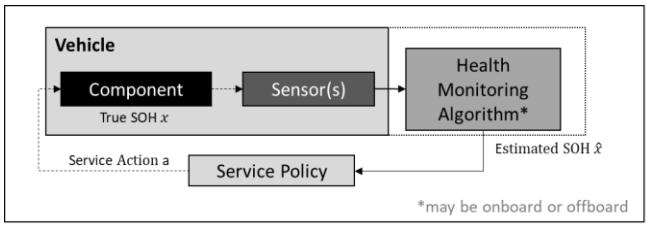

Figure 1: Generic predictive maintenance system**.**

There is a cost-benefit tradeoff for each realization of the HMA. Accurate RUL estimates improve maintenance planning, enable pre-ordering needed parts, and provide information needed to bundle maintenance to improve efficiency. For example, if a brake pad needs regularly scheduled servicing and an HMA indicates that the RUL of the brake rotor is shorter than the time to the next brake pad servicing, then replacing the rotor now will prevent an additional trip to service in the future. Developing an HMA to predict RUL is more complicated than developing a simple HMA to output a Boolean SOH, so the effort should only be undertaken if there is evidence that the additional savings will be worth the effort.

2.2. Valuation Framework Overview

The proposed framework assembles models of component degradation, HMA performance, and a policy of service actions taken in response to the HMA outputs. To simplify calculations, the framework is applied to discrete times throughout the vehicle life. At each discrete step, the true health state is represented as a probability distribution on some space of health states. Then, from a model of health monitoring algorithm performance, we get a probability distribution of the outputs from the algorithm given the ground-truth health state. Each output from the HMA is associated with a service action, as defined by some deterministic policy. Each of these actions has a cost and an effect on the health state of the component following the action. The overall cost of a maintenance strategy can then be assessed by calculating the expected cost from the assembly of models.

Formally, we define the valuation framework by the following five components, each of which is described in more detail in the following sections.

1. Vehicle Life Discretization
2. Degradation Model
3. Health Monitoring Algorithm Performance Model
4. Policy of Actions
5. Model of Action Outcomes

2.2.1. Vehicle Life Discretization

First, we must define the behavior of the PMS as it operates throughout the vehicle's life. This is captured by the *Vehicle Life Discretization*, in which we will assume that service actions are only taken when there is a new output from the health monitoring algorithm, and that these outputs come at discrete times in the set $\mathbb{L} = \{0, L_1, L_2, \dots, L_{design}\}$. Here, L_{design} is the ultimate lifetime of the vehicle, and the time by which we are aiming to optimize maintenance costs. Although we will use the terminology "time" for simplicity, the units of life may be any measure of vehicle use such as time, mileage, or number of rotations. We will separately denote the component life at step i by $L_i^c \in \mathbb{L}$. The difference between these lives is that the component life L_i^c may be re-set to 0 whenever the component is replaced, but the vehicle life L_i is strictly increasing with use.

2.2.2. Degradation Model

The *Degradation Model* quantifies the expected degradation of the component as a function of component life. First, we must define a state-space for the component ground-truth state of health, denoted $\mathbb{X}_{GT}$. This state-space could be a binary set (e.g. {healthy, faulty}), some larger discrete set (e.g. {100%, 90%, …, 0%}), or a continuum (e.g. [100, 0]). The choice of state space depends on the ability to accurately measure and model the true health state of the component.

With the state-space defined, this model shall define a conditional probability mass function of the state at each step $L_i \in \mathbb{L}$ given the history of all previous states.

$$P_{GT}(x_i \mid x_1, \dots, x_{i-1}, L_i^c),\ x_j \in \mathbb{X}_{GT},\ L_i^c \in \mathbb{L} \qquad (1)$$

In Expression (1), the notation $P(a|b)$ denotes the classic definition of conditional probability. This is a generic model that affords the designer of the valuation framework flexibility in modelling the degradation of the component. In many cases, a Markov assumption will hold and the dependence on any state older than x_{i-1} may be removed. Common reliability models may be expressed in this format, such as the Weibull distribution which is widely used in reliability modelling (Crowder, 1991). The Weibull distribution is defined by the cumulative distribution function (CDF) in Eq. (2) below.

$$F(t;\beta,\eta,\tau) = \begin{cases} 1 - \exp\left(-\left(\frac{t-\tau}{\eta}\right)^{\beta}\right); & t \geq \tau \\ 0\ ; & t < \tau \end{cases} \qquad (2)$$

This is known as the three-parameter Weibull distribution, which gives the cumulative probability that a component has failed by life t. The location parameter, τ, defines the minimum life that is expected to be failure free. Note that the two-parameter Weibull distribution is the specific case where τ is zero. The scale parameter, η, defines the relative scale of the distribution. The shape parameter, β, specifies the shape of the distribution. Both η and β are dimensionless, and τ is

in the same life units at t (hours, miles, revolutions, etc.). More advanced Weibull-derived reliability models may be applied, such as extended, exponentiated, truncated, or mixture Weibull models (Murthy, Bulmer, & Eccleston, 2004) (Xie & Lai, 1995).

For any component and model, Weibull parameters can be determined by conducting run-to-failure testing on a statistically significant set of parts and fitting the above distribution to the results. Goodness of fit tests can be used in conjunction with the principal of maximum likelihood to select an appropriate model. Some sources such as this table from GE (General Electric, 2018) can be used to estimate typical shape parameter values for different types of components.

Another realization of this model of ground-truth health state could employ failure records, such as warranty data, that capture the total number of parts failed within some population. A common warranty metric used in automotive is instances per thousand vehicles (IPTV) at 12, 24, 36, and 60 months operation. This data can be used to derive a P_{GT} distribution via interpolation.

2.2.3. Health Monitoring Algorithm Performance Model

The *Health Monitoring Algorithm Performance Model* defines the performance of the health monitoring algorithm as a function of the ground-truth health state of the component. First, we define the space of outputs of the algorithm, denoted by $\mathbb{X}_O$. It is not necessarily true that $\mathbb{X}_{GT}$ and $\mathbb{X}_O$ are equivalent, though it should be that the size of $\mathbb{X}_O$ is less than or equal to the size of $\mathbb{X}_{GT}$, as it would not be possible to design an algorithm to output more health states than can be accurately measured or modelled.

The simplest case is a Boolean health monitoring algorithm, with $\mathbb{X}_O = \{healthy, faulty\}$. More complex cases may include a state-of-health estimation algorithm that returns component health on a continuum from perfectly healthy to unusably faulty, or a prognostic algorithm that returns estimates of both the current SOH and the RUL. Note that $\mathbb{X}_O$ may be multi-dimensional to accommodate multi-output realizations of the HMA.

Second, we must define a model of the performance of the HMA conditioned on the true state of the component. We will denote this model $P_{HMA}(\hat{x}\,|x_e)$, where $\hat{x}$ is the output of the health monitoring algorithm and x_e is the expected output given the ground-truth health state of the component. In cases where $\mathbb{X}_O = \mathbb{X}_{GT}$, then x_e is equal to the ground-truth SOH (x). In other cases, x_e is the ideal output of the health-monitoring algorithm, as may be defined by some deterministic function $f_{HMA}(x)$.

If the state-space $\mathbb{X}_O$ is discrete, then the performance of the HMA can be captured by a confusion matrix (Stehman, 1997). For a classifier with N classes, the confusion matrix is defined by Eqn. 3.

$$CM = \begin{bmatrix} P(\hat{x} = c_1\,|x_e = c_1) & \cdots & P(\hat{x} = c_1\,|x_e = c_N) \\ \vdots & \ddots & \vdots \\ (\hat{x} = c_N\,|x_e = c_1) & \cdots & P(\hat{x} = c_N\,|x_e = c_N) \end{bmatrix} \quad (3)$$

where x_e is the expected output of the classifier, $\hat{x}$ is the class predicted by the classifier, and $\{c_1, \dots, c_N\}$ is the set of N classes that the sample may belong to. Note that the sum of each column of the confusion matrix must be equal to one for these probabilities to be well defined.

Alternatively, if the HMA outputs a predicted SOH in a continuous $\mathbb{X}_O$, then the outputs of the algorithm may be modelled as an error distribution. Experimental results should drive the choice of model here. In the simplest case, this could be a zero-mean gaussian distribution centered at the expected output. More complex models may include additional parameters to account for the state-space bounds, bias, heteroskedacity, skewness, multi-modal distributions, or other models that best fit a set of experimental results. Note, for generality, that $\mathbb{X}_O$ may be multi-dimensional, and the error model may account for correlation between dimensions.

It should be no surprise that the value of a PMS is inextricably linked to the performance of the health monitoring algorithm. The maximum value will come from an algorithm with perfect performance (100% accuracy and precision in the Boolean case, 0 error in the SOH/RUL estimate case), and negative value could be incurred from an algorithm with a high error rate. This model of HMA performance presents a degree of freedom in the valuation framework which can enable some deeper insights. For example, the framework can be used to determine the minimum acceptable HMA performance to meet a target ROI, or to tune the calibration of an existing HMA to maximize value.

2.2.4. Policy of Actions

The *Policy of Actions* defines the service actions taken when an output is issued by the HMA. This is modelled by a deterministic function f_a that returns a recommended action a_i from a set of actions $\mathbb{A}$, depending on the HMA output $\hat{x}_i$, and the component life L_i^c.

$$a_i = f_a(\hat{x}_i, L_i^c) \quad (4)$$

Some examples of actions, a_i, that a service policy may issue include:

- No Replacement: Do not replace component if $\hat{x}_i$ is healthy.
- Blind Replacement/Repair: Replace/repair component if $\hat{x}_i$ indicates a fault.
- RUL Scheduling: If $\hat{x}_i$ includes an RUL estimate, and the remaining life is less than the planned time before the next visit to the service hub, replace the component now.
- Inspect & Replace/Repair: If $\hat{x}_i$ indicates a possible fault, first inspect the component, and then only replace/repair it if the inspection reveals a fault.

- Failure-Free Period: Ignore any fault indicated by $\hat{x}_i$ if component life L_i^c is below some minimum.
- Maximum Life: Replace the component regardless of output state $\hat{x}_i$ if component life L_i^c is above some maximum.

The service policy dictates how the outputs of the HMA are consumed to realize its benefits. This valuation framework can be used as a tool for determining the optimal policy given an HMA of known performance. The optimal policy will depend on the failure rate of the component and the performance of the HMA. For example, it is not worth incurring any cost to inspect a component before replacing it if the HMA has perfect performance. In cases where the HMA has low fault detection performance, instituting a maximum life can help reduce overall cost.

2.2.5. Model of Action Outcomes

The final piece of the valuation framework is the *Model of Action Outcomes*. For each action in $\mathbb{A}$, we require a model of two outcomes: the cost, and the resulting SOH of the component being monitored.

$$P_{cost}(c_i \mid a_i, x_i), \qquad c_i \in \mathbb{R}^+, a_i \in \mathbb{A}, x_i \in \mathbb{X}_{GT}$$
$$P_{service}(x_i' \mid a_i, x_i), \qquad a_i \in \mathbb{A},\ x_i, x_i' \in \mathbb{X}_{GT}$$

The cost model presented above is the most generic form, allowing for both dependence on the true health state x_i and probabilistic definition. Some actions will have a deterministic cost, regardless of the true ground-truth health state. For example, if the action is blind replacement if $\hat{x}_i$ indicates a fault, then the cost is the cost of the replacement regardless of the ground truth. If the action is to first inspect the component, and replace it only if needed, then the cost is the inspection only if x_i is healthy, or the cost of inspection plus replacement if x_i is faulty. Probabilistic modelling can capture costs that are variable in nature, such as towing which may be a function of distance and time of day.

Finally, note that the costs may include both financial expenses as well as soft-costs, such as customer safety and comfort. Quantifying soft costs can be difficult and may require assumptions based on market research. If a dollar-equivalent of each soft-cost can be derived, then the inputs of the framework can be optimized to minimize the overall cost in a single objective optimization problem. If such an equivalence is not possible, then the framework can be implemented to output multiple costs. For example, the costs in an automotive setting may include total maintenance cost plus the cost of customer dissatisfaction in the event of a roadside failure. This may manifest in a challenging multi-objective optimization problem with conflicting objectives, as increasing maintenance costs will likely decrease dissatisfaction, and vice versa.

The service model, $P_{service}$, recognizes the fact that service actions may not be perfect. If the action is to repair a part and there is some risk that the repair fails, that can be captured in this distribution. If the action is to replace the component, and there is no risk of the replacement failing, then $P_{service}$ can be collapsed to an elementary event which guarantees x_i' is completely healthy.

2.3. Evaluating the Framework

The components of the predictive maintenance system valuation framework are summarized below in Table 1.

Table 1: Summary of valuation framework components.

Component	Sub-Component	Notation
Design Life Discretization	System design life	L_{design}
	Life discretization	$\mathbb{L} = \{0, L_1, \dots, L_{design}\}$
Degradation Model	Ground-truth space	$\mathbb{X}_{GT}$
	Model $\mathbb{L} \times \mathbb{X}_{GT}^{i-1} \rightarrow \mathbb{X}_{GT}$	$P_{GT}(x_i \mid x_1, \dots, x_{i-1}, L_i)$
Health Monitoring Algorithm Performance Model	Output space	$\mathbb{X}_O$
	Model $\mathbb{X}_O \rightarrow \mathbb{X}_O$	$P_{\mathrm{HMA}}(\hat{x} \mid x_e)$
Policy of Actions	Action space	$\mathbb{A}$
	Function $\mathbb{X}_O \times \mathbb{L} \rightarrow \mathbb{A}$	$f_a(\hat{x}_i, L_i^c)$
Model of Action Outcomes	Model $\mathbb{A} \times \mathbb{X}_{GT} \rightarrow \mathbb{R}^+$	$P_{cost}(c_i \mid a_i, x_i)$
	Model $\mathbb{A} \times \mathbb{X}_{GT} \rightarrow \mathbb{X}_{GT}$	$P_{\mathrm{service}}(x_i' \mid a_i, x_i)$

Finally, we need a process to combine the components of the valuation framework to determine the expected cost of the PMS. In even the simplest formulations of this framework, calculating the expected value directly will prove to be a complicated task. Challenging integrals may arise when calculating expectations of the multiple PDFs employed in this framework, and solutions may not be generalizable to slight variations in the formulation. Instead of relying on direct computation, it is desirable to define a generic estimation method that can be applied to any formulation of the valuation framework described above.

One suitable approach would be to implement a Monte-Carlo simulation. Each iteration of the simulation would begin with a sample brand-new vehicle with brand-new components. Then, with each iteration i of the predictive maintenance system throughout the vehicle life discretization $\mathbb{L}$, the simulation would sample a representative SOH, x_i, from the degradation model, P_{GT}. The output of the HMA, $\hat{x}_i$, given this SOH is sampled from the conditional model of HMA performance, $P_{HMA}(\hat{x}_i \mid x_i)$. The action to be taken is determined from the policy of actions, $f_a(\hat{x}_i,\ L_i^c)$, and the cost of this action is then sampled from $P_{cost}(c_i \mid a_i, x_i)$. The resulting component SOH is updated according to $P_{service}(x_i' \mid a_i)$, and the simulation repeats until the design

life of the vehicle is met. This process is summarized in Figure 2.

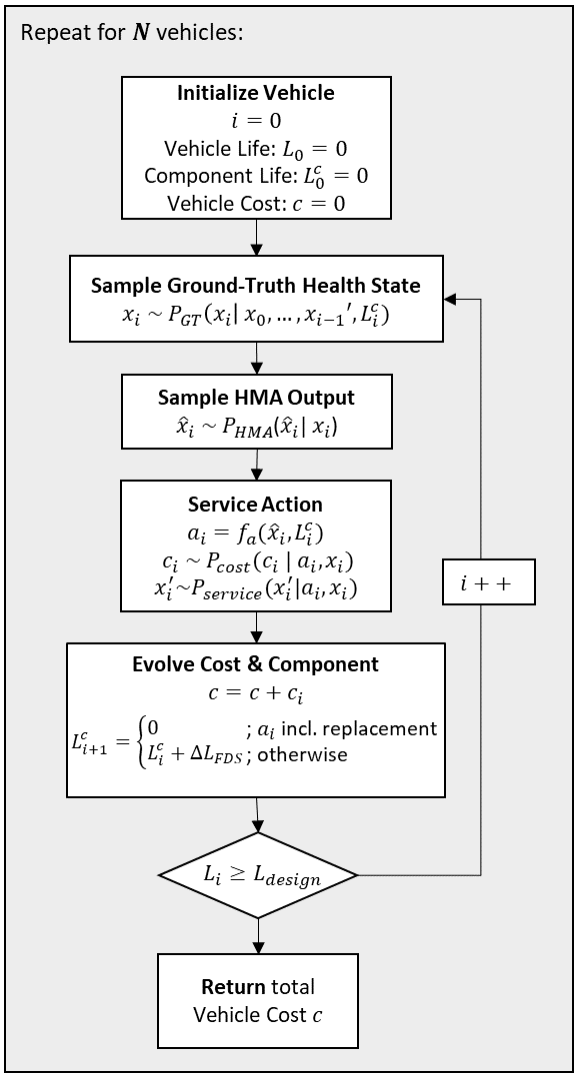

Figure 2: Monte Carlo simulation block-diagram.

This simulation and sampling process is repeated for a large amount (N) of simulated vehicles, yielding a distribution of costs for the maintenance strategy and enabling calculation of the expected cost. Ultimately, this Monte-Carlo simulation can be used to valuate a predictive maintenance system by comparing the expected costs from a Policy of Actions that uses information from an HMA to a Policy of Actions that does not.

3. RESULTS & DISCUSSION

3.1. Applications

There are many ways this framework can be employed to yield valuable insights, and it can serve as a useful tool throughout the project cycle.

3.1.1. Project Valuation

When looking to define a project to develop a predictive maintenance system and prove that the project will yield value, this framework can be employed to estimate a dollar value for the project. This is achieved by comparing the expected cost of a service strategy that does not include a HMA against a service strategy that does. Given an estimated cost to develop and employ the PMS that accounts for engineering effort, testing materials, and any additional sensors or processors required to implement the HMA, this valuation can be used to derive the ROI of the effort. An honest assessment will present the valuation as a range, in which the maximum value is attained with a perfect HMA ($P_{HMA}(\hat{x}|x_e) = 1_{\hat{x}=x_e}$), and the expected value considers some non-perfect performance that the development team believes is attainable.

3.1.2. Requirement Derivation

In some scenarios, engineering teams may be required to meet a minimum acceptable ROI for their efforts to develop a new maintenance system. Given this requirement, it can be possible to derive minimum performance requirements for the health monitoring algorithm. Consider, for example, a binary HMA. Savings come from true positives that allow for service actions to prevent in-use failures resulting in costly downtime. False positives, on the other hand, can drive up cost from trusting an HMA that isn't accurate. Therefore, given a minimum acceptable ROI, this framework may be used to derive the minimum true positive rate (TPR) and maximum false positive rate (FPR) for the HMA. These requirements may then be used as acceptance criteria after the R&D stage in which the algorithm is developed and its performance known.

3.1.3. Service Policy Optimization

After an HMA is developed with known performance, this framework can be used to optimize the service policy. For example, if the HMA has a relatively high rate of false detections, then the policy should be revised to inspect components before they are replaced (versus blind replacement, which will result in unnecessary part cost). However, depending on the component, this inspection cost may be too high to justify. By employing the framework to a variety of scenarios, service decision makers can make data-driven decisions to minimize their costs.

3.1.4. Calibration Tuning

Health monitoring algorithms often require many calibratable parameters. These parameters control the performance of the algorithm, and there are often conflicting objectives. Consider a case where an indicator signal is compared to a threshold, and any value above that threshold is labelled as a fault. Increasing the threshold will reduce the risk of false positives, but also reduce the rate of true positives. This

tradeoff relationship is captured by a receiver operating characteristic (ROC) curve. Once the ROC curve of the HMA is known, the valuation framework can be used to identify the set of calibrations that optimizes cost savings.

3.2. Example

To demonstrate the benefits of this generalized framework and its potential uses, we have constructed a simple example. Suppose we have a component where inspection cost is an order of magnitude below replacement cost, which itself is an order of magnitude below the cost of an in-service failure. For simplicity, we will use the values in Table 2.

Table 2: Example cost parameters.

Action	Cost
Inspect ($c_{inspect}$)	10
Replace ($c_{replace}$)	100
In-Service Failure (c_{fail})	1000

We will consider a simple binary diagnostic HMA implemented on this component, such that the output is either "healthy" or "faulty". We will consider a design life (L_{design}) of 100,000 miles, with a discretization at 1 mile steps (assuming the HMA executes once per mile of driving). The degradation model will be a simple two-parameter Weibull distribution, given by Eqn. 2 with τ equal to zero. This formulation gives that both $\mathbb{X}_{GT}$ and $\mathbb{X}_O$ are the binary set {healthy, faulty}.

For this analysis, we will compare 6 different policies of actions, defined below.

1. Corrective (f_C): only replace the component if an in-service failure occurs.
2. Scheduled (f_S): replace the component if an in-service failure occurs, or a pre-defined service life L_s is reached. L_s will be defined as an integer division of the design life (i.e. there will be N replacements at intervals of $L_{design}/(N+1)$ miles).
3. Blind Replacement (f_R): replace the component if an in-service failure occurs, or the HMA issues a "faulty" output.
4. Inspect & Replace (f_{IR}): replace the component if an in-service failure occurs. If the HMA issues a "faulty" output, first inspect the component, and replace it if the inspection confirms the fault.
5. Scheduled x Blind Replacement (f_{SR}): replace the component if an in-service failure occurs, the service life L_s is reached, or the HMA issues a "faulty" output.
6. Scheduled x Inspect & Replace (f_{SIR}): replace the component if an in-service failure occurs, the service life L_s is reached, or the HMA issues a "faulty" output and an inspection confirms the fault.

For the model of action outcomes, we will assume that both replacement and inspection actions have 100% probability of success, and that the costs are deterministic as defined in Table 2.

The main question to be asked by the health monitoring development team is whether there is a business case to justify developing an HMA for this component. This will depend on the expected failure rate of the component, which we are assuming may be modelled by a two-parameter Weibull distribution with shape parameter β and scale parameter η.

First, we must determine the "reference" maintenance policy for comparison. This will be the lowest cost policy that does not consume any outputs from an HMA – in this case, either a corrective (f_C) or scheduled (f_S) policy. Figure 3 below shows which reference policy is best as a function of Weibull shape and scale parameter.

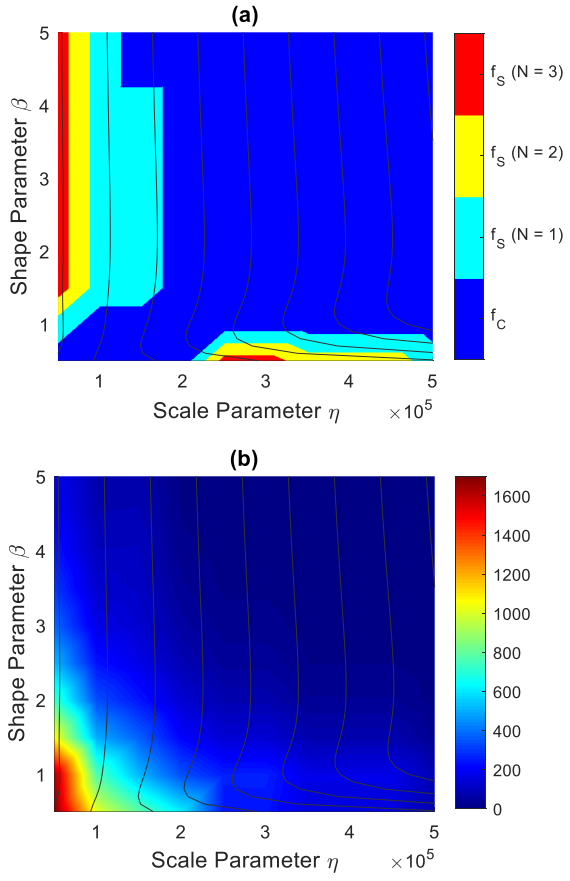

Figure 3: (a) Shows the reference policy with minimum cost, (b) shows the savings of an HMA policy versus the optimal reference strategy as a function of the component's Weibull parameters.

The lines in Figure 3 are sets of shape and scale parameters with equivalent expected life. It can be seen in Figure 3 (a) that for shape parameters greater than 1, the optimal number of scheduled replacements is strongly anti-correlated with the scale parameter. It is interesting to note that for components with high infant mortality rates ($\beta < 1$), a scheduled replacement strategy can actually drive up costs by replacing matured components with new ones that are more susceptible to failure. For these components the optimal reference policy is simple corrective maintenance until the scale parameter is large enough.

Figure 3(b) shows the maximum possible savings that an HMA policy can yield versus the optimal reference policy. The savings are presented in expected dollars per vehicle saved. Not surprisingly, the expected savings increases as the scale and shape parameters decrease. It is intuitive that developing an HMA will be most valuable for components with high failure rates. These results are highly dependent on the cost parameters in Table **2** 2, and would differ significantly if c_{fail} and $c_{replace}$ were to change. The benefit of an HMA policy is in correcting field failures before they occur. Therefore, the larger c_{fail} is relative to $c_{replace}$, the more valuable an HMA will be. Correctly setting these costs is crucial to deriving correct insights from this valuation framework.

This analysis enables initial decision making if the value of developing an HMA is worth the effort. Given the estimated Weibull parameter for the component under study, the expected value of the project can be assessed using the heat map in Figure 3 (b), and the ROI can be estimated. Note that uncertainty in the Weibull parameters should be incorporated, as Figure 3 (b) shows that the value of an HMA is sensitive to the degradation model.

The above analysis assumes that an HMA with perfect performance is developed, which is likely not the case. HMAs with false negatives (missed detections) will lead to costly in-service failures, and HMAs with false positives (incorrect detections) will lead to unnecessary repair actions. We capture these imperfections in the framework by specifying the performance model P_{HMA}. For a binary $\mathbb{X}_{GT}$ and $\mathbb{X}_O$, this model is uniquely defined by two parameters: the false positive rate (FPR) and true positive rate (TPR), given by Eqn. 5.

$$\text{FPR} = P_{HMA}(\hat{x} = faulty \mid x_e = healthy) \quad (5a)$$

$$\text{TPR} = P_{HMA}(\hat{x} = faulty \mid x_e = faulty) \quad (5b)$$

Figure 4 (a) shows the optimal service policy as a function of HMA performance for a component with Weibull parameters $\beta = 1.3$, $\eta = 5 * L_{design} = 500{,}000$ miles. It is expected that about 10% of these parts will fail within the 100,000 mile vehicle life.

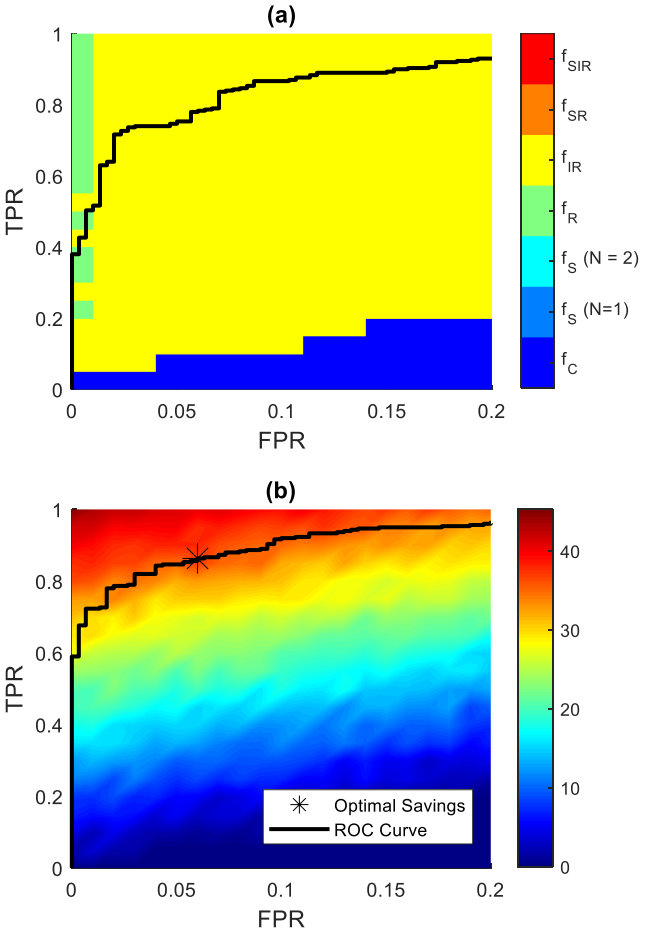

Figure 4: (a) the optimal service policy as a function of HMA performance. (b) the expected cost savings of the optimal policy, in $ per vehicle. The black line on each plot shows the ROC curve of a sample HMA.

This analysis quantifies some intuitive behaviors. We can see that "blind replacement" is only the optimal policy when the TPR is high and FPR is low, when we have full trust in the HMA. The dominant optimal policy is f_{IR}, which is not surprising given the relatively low cost of inspection vs replacement. If the cost of inspection were to increase relative to replacement, we would expect the region where f_R is optimal to grow to accept greater false positive rates. Finally, the blue region captures all TPR/FPR performance metrics where the HMA will result in negative value and sticking with reactive maintenance is the best strategy. For this formulation of the PMS, f_{SR} and f_{SIR} are never the optimal policies. It is expected that scheduling becomes an important component of the maintenance policy to protect against false negatives when the failure rate is higher and the TPR is low.

If there is a minimum ROI expected from the development, this analysis can be used to derive the performance requirements for the HMA by identifying the region on Figure 4 (b) that yields acceptable ROI. This analysis can also be used for calibration tuning once the HMA performance is

known. Suppose, for example, the HMA performance is as defined by the ROC curve shown in black on Figure 4. Any point on this performance curve is attainable by changing the algorithm calibrations. The valuation framework can be used to identify the point on the ROC curve with maximum savings, as shown by the marker on Figure 4 (b).

In this simple example, we have shown how the valuation framework can be used to valuate projects, optimize maintenance policies, set HMA performance requirements, and derive cost-optimal calibrations. The tools and analysis outlined in this simple example may be expanded to more advanced degradation models, HMA models (i.e. SOH and/or RUL estimation algos), and service policies.

4. CONCLUSION

The framework presented in this paper aims to present a comprehensive and flexible method for estimating the costs associated with a maintenance policy that incorporates a health monitoring algorithm. This contributes to the existing literature on maintenance cost modelling and optimization by standardizing model components needed. The presented example shows how this framework may be employed to derive insights and drive decisions regarding investment in developing HMAs. It was also discussed how this framework can deliver insights deeper than simple project valuations, such as deriving HMA performance requirements, tuning HMA calibrations, and optimizing service policies. OEMs can use a framework like this to aid with deciding where to spend engineering effort as they work to develop health-aware vehicles and minimize maintenance costs for their customers.

REFERENCES

Advanced Technology Services, Inc. (2020). *Transforming From PM To PDM: Why Factories Are Making The Switch.* Advanced Technology Services, Inc.

Alrabghi, A., & Tiwari, A. (2013). A Review of Simulation-based Optimisation in Maintenance Operations. *UKSim 15th International Conference on Computer Modelling and Simulation*.

Cal. Code of Regulations. (2021). California Code of Regulations, Title 13, Division 3.

Crowder, M. J. (1991). *Statistical analysis of reliability data.* Chapman & Hall.

Decaix, G., Gentzel, M., Luse, A., Neise, P., & Thibert, J. (2021, 04 03). *A smarter way to digitize maintenance and reliability*. (McKinsey) Retrieved 06 09, 2021, from https://www.mckinsey.com/business-functions/operations/our-insights/a-smarter-way-to-digitize-maintenance-and-reliability#

Deloitte Analytics Institute. (2017). *Predictive Maintenance: Taking pro-active measures based on advanced data analytics to predict and avoid machine failure.* Deloitte Consulting.

Fuchs, S., Safar, S., & Kok, E. (2016, 12 13). *Smaller fleet, big impact*. Retrieved from McKinsey & Company: https://www.mckinsey.com/business-functions/operations/our-insights/smaller-fleet-big-impact

General Electric. (2018). *About Weibull Distribution*. (General Electric Company) Retrieved 06 04, 2021, from https://www.ge.com/digital/documentation/meridium/Help/V43050/Default/Subsystems/ReliabilityAnalytics/Content/WeibullDistrubution.htm

Jahangirian, M., Eldabi, T., Aisha, N., Stergioulas, L., & Young, T. (2010). Simulation in manufacturing and business: A review. *European Journal of Operational Research*, 1-13.

Jonge, B., & Scarf, P. (2019). A review on maintenance optimization. *European Journal of Operational Research*.

Khandelwal, D. N., Sharma, J., & Ray, L. M. (1979). Optimal Periodic Maintenance Policy for Machines Subject to Deterioration and Random Breakdown. *IEEE Transactions on Reliability*.

Murthy, P., Bulmer, M., & Eccleston, J. A. (2004). Weibull model selection for reliability modelling. *Reliability Engineer & System Safety, 86*(3), 257-267.

ProAxion Inc. (2021). *Calculating the Impact of Predictive Maintenance on Productivity*. (ProAxion Inc.) Retrieved 06 09, 2021, from https://www.proaxion.io/calculate-predictive-maintenance-impact

Prometheus Group. (2020). *Five Areas to Consider Before Switching to Predictive Maintenance*. (Prometheus Group) Retrieved 06 09, 2021, from https://www.prometheusgroup.com/posts/five-areas-to-consider-before-switching-to-predictive-maintenance

SAE. (2012, 02 23). E/E Diagnostic Test Modes J1979_201202. Society of Automotive Engineers International.

SAE. (2013, 03 07). Diagnostic Trouble Code Definitions J2012_201303. Society of Automotive Engineers International.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 77-89.

Uptake. (2021). *How We Measure the Value of Predictive Maintenance for Truck Fleets*. (Uptake) Retrieved 06 09, 2021, from https://www.uptake.com/blog/how-we-measure-the-value-of-predictive-maintenance-for-truck-fleets

Xiang, Y., Cassady, C. R., & Pohl, E. A. (2012). Optimal maintenance policies for systems subject to a Markovian operating environment. *Computers & Industrial Engineering*, 190-197.

Xie, M., & Lai, C. D. (1995). Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function. *Reliability Engineering and System Safety, 52*, 87-93.

BIOGRAPHIES

Graeme Garner received the B.S.E. from Queen's University in 2018, where he studied a dual program of Applied Mathematics and Mechanical Engineering. His academic achievements include winning the J.B. Stirling Gold Medal for graduating with the highest academic standing in his class and receiving the Keyser prize for his research project on adapting Q-learning to decentralized stochastic control problems. He has a diverse professional background, including developing oil production forecasting algorithms at the Alberta Energy Regulator and studying automated trading strategies at the Canadian Imperial Bank of Commerce. He is currently at the Canadian Technical Center of General Motors, where he develops prognostics algorithms for vehicle hardware. His research interests are in robotics and intelligent systems.

Paola Santanna works at General Motors, where she is the design release engineer for gears and shafts.

Hossein Sadjadi received his Ph.D. degree in electrical engineering from Queen's University, Canada, and M.Sc. degree in mechatronics and B.Sc. degree in electrical engineering from the American University of Sharjah, UAE. He has been working at General Motors, Canadian Technical Center, Markham, ON, since 2017, and is currently the Global Technical Specialist for Vehicle Health Management. He also has served as a post-doctoral medical robotic researcher at Queen's university, senior automation engineer for industrial Siemens SCADA/DCS solutions, and senior mechatronics specialist at AUS mechatronics center. His research interests include autonomous systems and medical robotics. He has published numerous patents and articles in these areas, featured at IEEE transactions journals, and received several awards.