Abstract—Numerous prognostic methods have been developed, aiming at predicting future system reliability with the highest possible accuracy. It is striking that the relation with the subsequent maintenance optimization process is generally overlooked, while it is important in practice. Additionally, almost all existing methods are based on a single degradation measure, and focus on systems with only one degradation and failure mode. In practice, however, multiple degradation measures are often available and needed to adequately predict future system degradation. Moreover, systems may suffer from various kinds of faults, all resulting in different degradation behaviors. To accommodate these properties, we establish a link between failure prognosis and maintenance optimization, and accordingly propose a multivariate multiple-model approach to system reliability prediction. We conclude that in the presence of multiple degradation modes and provided they are correctly identified, a multiple-model approach outperforms a single-model approach with respect to prediction accuracy. Moreover, in the presence of multiple degradations and failure modes, overall predictions of the remaining useful life as generated by common prognostic approaches are not directly suited for maintenance decision making, as different kinds of system failures and maintenance activities are associated with different costs. In contrast, our approach yields conditional predictions of future system reliability, which much better suit the maintenance optimization process.

Index Terms—Reliability prediction; condition-based maintenance; degradation modeling; prognostics; multivariate time series analysis; recursive Bayesian filtering.

I. INTRODUCTION

CONDITION-BASED maintenance is an increasingly popular maintenance strategy in which maintenance activities are planned based on information collected through real-time condition monitoring. Its promise is twofold [1]: first, unnecessary maintenance can be eliminated, reducing maintenance costs; second, failures can be avoided, improving safety and reducing unscheduled downtime.

Condition-based maintenance comprises:

1) fault diagnosis;
2) failure prognosis;
3) maintenance planning.

Fault diagnosis concerns the detection of faulty behavior and the determination of its cause(s). Prognosis refers to the prediction of future degradation behavior and the estimation of the associated failure time. Finally, maintenance planning comprises the determination of the optimal time and type of maintenance. The focus of this work is on prognostics, i.e. modeling and forecasting of degradation behavior.

Although in recent years a lot of attention has been devoted to failure prognosis (see Section II), failure prognosis is still an emerging research area with a number of open challenges [3]–[5]. In this paper, we address two of them. The first challenge is to establish a link between failure prognosis and maintenance optimization [3], [5]. Most existing prognostic methods have been developed without explicitly considering how the method is going to be used for maintenance optimization [5]. Accordingly, most existing methods for condition-based maintenance optimization base their maintenance decisions solely on diagnostic information, without considering prognostic information [6]. Although the link to the subsequent maintenance planning process is often overlooked, it is important in practice [3], [5]. The second challenge is the development of methods that can deal with multiple degradation modes and multiple degradation measures [4]. Most existing methods are based on a single degradation measure and account for just one degradation mode. In practice, multiple degradation measures are often available. Moreover, systems may suffer from various kinds of faults, all resulting in different degradation behaviors. Therefore, improvement in prediction accuracy is expected when considering multiple degradation measures and accounting for variability due to different fault causes.

To address the aforementioned challenges, we propose a multivariate multiple-model approach to degradation forecasting and reliability prediction. Multiple models are considered because degradation can be caused by various system faults. In general, for different system faults, system degradation evolves differently over time. So, to adequately model degradation behavior in the presence of multiple degradation modes, a distinct (multivariate) degradation model is needed for each degradation mode. Based on these models the future system reliability conditional to each degradation mode can be predicted.

We consider a multivariate approach because in practice often multiple degradation measures are available and required to adequately model the degradation process. In the case of a single degradation measure, system reliability and remaining useful life follow from comparing the predicted degradation signal with a predefined threshold value. In the case of multiple degradation measures, the computations of the system reliability and remaining useful life are less straightforward. In this work, we consider the computation of the system reliability in the multivariate case for various types of relationships among degradation measures.

For the degradation forecasting, a choice needs to be
made between a model-based [3], a model-free (artificial intelligence) [3], [7], or a hybrid (e.g. statistical) [3], [4] approach. We consider statistical approaches since they are particularly suited to manage and represent the uncertainty inherent to degradation forecasting. Good management and representation of uncertainty is of paramount importance for the subsequent decision making process. More specifically, we consider stochastic state space models, which can include most common uncertainty sources inherent to the forecasting process, i.e., temporal uncertainty, case-to-case (or sampling) variability, and measurement uncertainty [8], [9]. In addition, Bayesian filtering and prediction are used to estimate and account for the added value of using a multiple-model approach over a single-model approach. Finally, in Section IX conclusions and possible directions for further research are given.

II. RELATED WORK

Over the past few years, various prognostic methods have been proposed, ranging from model-based approaches to artificial neural networks and stochastic filtering approaches. Overviews of the various prognostic methods can be found in the review papers [3], [4], [7], [10].

Especially statistical approaches have received a lot of attention in the literature thanks to their ability to handle the uncertainty inherent to the degradation forecasting process. For instance, (hidden) Markov models [11]–[13] and models based on gamma [14], [15] and Wiener processes [9], [16], [17] have frequently been proposed for prognostic purposes. Nevertheless, most of the existing methods take only part of the uncertainty into account. For example, the authors of [17], [18] omit measurement variability, while the authors of [19], [20] consider measurement variability, but omit the case-to-case or temporal variability. Recently, in [9], [21] methods have been proposed that take both measurement uncertainty, temporal variability, and case-to-case variability into account. In this paper, we extend these methods to the multivariate case. Moreover, we address modeling uncertainty by considering a multiple-model approach. In contrast to the authors of [22], who consider multiple models to account for different stages in the degradation process, we consider multiple models to account for different qualitative degradation behaviors associated with different fault causes. The main difference compared to the methods proposed in [13], [23], [24] is that in [13], [23], [24] multiple hidden Markov models are considered to account for different rates of degradation, while we consider multiple parametric models to describe fault-specific degradation behavior. Variations in degradation rates are accounted for in stochastic model parameters, which are updated online based on observed monitoring data.

Almost all existing prognostic methods are based on a single degradation measure. An exception is the approach proposed in [25]. In this work, we extend the approach proposed in [25] to a multiple-model approach, taking into account different degradation modes. In contrast to the authors of [25], we consider a Wiener process-based degradation model accounting for sampling, measurement, and temporal variability. Moreover, we distinguish between different types of possible relationships among the degradation measures (e.g. between redundant and complementary measures).

Characteristic for most methods proposed on prognostics is that the link to the subsequent maintenance optimization process is missing. A small number of papers on maintenance optimization briefly consider the link to the prognosis process [6], [26]–[28]. These maintenance methods are however based on a single degradation measure and consider only one degradation mode. In this paper, we establish a link between diagnosis, prognosis, and maintenance optimization in the case of multivariate degradation measures and multiple degradation and failure modes.

III. PROBLEM FORMULATION

A. Terminology

In the sequel, the following terminology is used (see Figure 1). First, we made a distinction between three types of system behavior:

1) healthy behavior;
2) faulty behavior;
3) system failure.

Healthy behavior refers to the situation in which the system functions as desired. Faulty behavior describes the situation in which the system exhibits some aberrant behavior, but is still functional. When at least one of the system tasks can no longer be executed properly, we talk about a system failure. The transition from healthy behavior via faulty behavior to a system failure can take different forms, which we call degradation modes. So, the degradation modes $d_1$ till $d_r$...
T o handle case-to-case variability, methods have the use of a fixed model for degradation forecasting is unde-
(e.g. due to different environmental or operating conditions), behavior varies for different fault types and from case to case of the degradation measure(s) over time. As the degradation measures over time is characteristic for the type of fault or in a specific failure mode. The evolution of the degradation measures indicate to what extent the system is healthy, faulty, and fault types as follows: The values of the degradation measures are linked to the failure modes X deg%ation [1], [25]. The set of degradation measures is a continuous variable that can be computed from sensor information and that is highly correlated with system

C. Goals

The first step of failure prognosis comprises the forecasting of the degradation measure(s) over time. As the degradation behavior varies for different fault types and from case to case (e.g. due to different environmental or operating conditions), the use of a fixed model for degradation forecasting is undesired [1], [9]. To handle case-to-case variability, methods have been proposed that model the degradation by a parametric model with stochastic parameters [1], [9], [29]. Our first goal is to augment these single-model approaches to a multiple-model approach, where a distinct model is defined for each degradation mode. The aim is to explicitly model variability due to different fault types, so as to reduce modeling error.

As a third goal, we aim to explicitly make the link to the subsequent maintenance optimization process. Failure prognosis is not an isolated task, but a task within the condition-based maintenance process. We therefore analyze the dependencies between diagnosis, prognosis, and maintenance optimization. Moreover, we investigate how the prognostic result should be specified to support maintenance optimization in the case of multiple degradation and failure modes.

IV. MOTIVATING CASES

We motivate the need for multivariate prognostic methods accounting for multiple degradation and failure modes, based on two practical examples: failure prognosis for railway tracks and failure prognosis in buildings.

A. Failure prognosis for railway tracks

A key component of a railway network is the track. Besides that the track provides trains with a dependable surface for their wheels to roll on, it is an essential part of the train detection process using track circuits. For the purpose of train detection, at one end of each railway section, an electrical current is transmitted. In the absence of a train, this current flows via the rails to the other end of the section, where it is measured by a receiver. When the current measured at the receiver exceeds a certain threshold, the section is reported as free. When a train is present, the circuit is shorted by the wheels of the train and the current measured at the receiver is close to zero, in which case the section is reported as occupied (see Figure 2). To guarantee reliable train detection, it needs to be ensured that the conductance properties of the rails and the shunting properties of track and trains are of high quality.

Fig. 1. Relationships between fault types, degradation modes, and failure modes.

Fig. 2. Flow of current in a railway track circuit.
All together, the railway track serves the following purposes:

1) safe and comfortable guidance of trains;
2) correct detection of a free track;
3) correct detection of an occupied track.

Accordingly, three failure modes are defined (see Figure 3).

The proper execution of the aforementioned tasks may be impaired by different faults, four common ones of which are:

- $f_{rc}$: rail contamination;
- $f_{rd}$: rail surface defect;
- $f_{ed}$: electrical disturbance;
- $f_{ij}$: insulated joint defect.

The faults are related to the failure modes $c_1$ till $c_3$ as follows (see Figure 3): Contamination between the rail surface and the wheels (e.g. rust films, sand, and leaves) may hamper both the safe and comfortable guidance of trains and the correct detection of an occupied track (because the contamination hinders passing trains to shorten the circuit). Both rail surface defects and insulated joint defects may impair the safe and comfortable guidance of trains, as well as the correct detection of a free track. Finally, electrical disturbances may hamper the correct detection of a free track.

Faults $f_{rc}$, $f_{rd}$, $f_{ed}$, and $f_{ij}$ are all associated with different time behaviors of degradation (see Figure 3), where a distinction is made between the following three types of qualitative degradation behavior:

- $d_1$: linear;
- $d_2$: exponential;
- $d_3$: intermittent.

From the above description, it follows that adequate degradation modeling for railway tracks requires a multivariate multiple-model approach. First, the railway track is subject to different degradation modes. For example, a false positive train detection (i.e. failure mode $c_2$) can be caused by both a rail defect, an insulated joint defect, or an electrical disturbance. How the degradation evolves over time depends to a large extent on the type of fault present. Therefore, multiple models are required to forecast degradation behavior. Second, multiple degradation measures are needed to adequately forecast degradation behavior. The system’s ability to detect a free track is expressed in the current flowing through the track circuit receiver when the track is vacant [30]. The system’s ability to detect an occupied track is reflected in the current not flowing through the track circuit receiver when the track is occupied by a train [30]. Among other things, the vertical axle box accelerations [31], [32] provide information about the system’s ability to safely and comfortably guide vehicles. So, for this application it is not possible to adequately model all degradation behaviors using just one measure, e.g. the guidance abilities cannot be assessed adequately using just electrical information, whereas the detection abilities cannot be assessed adequately based on just mechanical information.

B. Failure prognosis in buildings

Heating, Ventilation, and Air Conditioning (HVAC) systems are another example of systems that fullfill multiple tasks and that are subject to different degradation modes. Without going into detail, an HVAC system serves the following purposes:

1) temperature control;
2) humidity control;
3) ventilation.

Accordingly three failure modes can be defined. Multiple faults can be identified that hinder the proper execution of one or more of these tasks. Just a few examples are [33]:

- $f_{mb}$: malfunctioning boiler;
- $f_{sv}$: stuck heating/cooling coil valve;
- $f_{di}$: deteriorating supply fan;
- $f_{dd}$: deteriorating damper (controlling the mixing ratio between outside and re-circulation air).

Like for the railway example, multiple degradation measures are needed to model degradation behavior; the system’s ability the control zone air temperature is expressed in zone temperature measures, while the system’s ability to regulate humidity is reflected in humidity (correlated) measures, and the ventilation quality is reflected in CO$_2$ (correlated) measures. Because some of the faults affect multiple system goals (e.g. a deteriorating supply fan may affect all goals) the degradation behavior of the different measures may be correlated. Therefore, it is advantageous to consider multivariate degradation modeling at the system level, rather than looking at the individual tasks.

V. Prognostics within the Condition-based Maintenance Process

Condition-based maintenance aims to optimize maintenance planning by making use of real-time monitoring data. The path from the monitoring data to an optimal maintenance decision includes data pre-processing, fault diagnosis, failure prognosis, and maintenance optimization. Besides the proper implementation of the individual processes, adequate incorporation of the dependencies between the individual processes is crucial for the success of condition-based maintenance. With respect to failure prognosis, the following dependencies are relevant:

1) dependencies between the diagnosis and prognosis processes;
2) dependencies between failure prognosis and maintenance optimization.

A. Dependencies between diagnosis and prognosis

Although fault diagnosis and failure prognosis concern identical tasks, and are often treated individually, exploiting their mutual dependence is valuable for both diagnosis and prognosis. As outlined before, different fault types are associated with different degradation behaviors. So, information regarding the type of fault present (diagnostic result) provides information about future degradation behavior. Vice versa, information about degradation trends (prognostic result) provides information regarding the type of fault present. We propose to exploit this dependence by using the diagnostic result to determine the likelihood of each prognostic (fault-specific) model.
B. Dependencies between prognosis and maintenance optimization

The prognostic result serves (together with the diagnostic result) as an input for the maintenance optimization process. It is therefore important to ensure that the prognostic result is specified such that it facilitates maintenance optimization. An adequate specification of the prognostic result requires an understanding of the maintenance optimization process. Therefore, before defining an adequate specification of the diagnostic result, background information on the maintenance optimization process is given.

1) Maintenance optimization process: Maintenance optimization is a typical example of a decision task subject to risk and uncertainty: we have uncertainty regarding the current and future system health, and consequently, we have the risk of making non-optimal maintenance decisions. In the presence of risk and uncertainty, decisions are commonly made based on the expected utility theory [34], which is a framework for determining the best (maintenance) decision given probabilistic information regarding the actual situation\(^2\). In the sequel, we assume that maintenance decision are made based on the expected utility theory (see Appendix A for a brief explanation of the expected utility theory).

In contrast to common maintenance optimization methods, which limit the maintenance optimization task to deciding whether or not to perform preventive maintenance at a particular time instant, we augment this task by deciding on the following items [35]:

1) the required type of maintenance;
2) the optimal time to perform maintenance.

So, the possible maintenance decisions are:

\[ d_{0,T} : \text{do nothing}; \]
\[ d_{a,t} : \text{perform maintenance activity } a \in A \text{ at time } t \in T, \]

with \( a \) and \( t \), in turn, decision variables, \( A \) the finite set of possible maintenance activities, and \( T \) the discrete set of available maintenance time instants. The goal is to find the combination of type \( a \) and time \( t \) of maintenance that minimizes the total maintenance costs \( C_{\text{total}}(a,t) \). The expected total maintenance costs can be computed as [35]:

\[
E(C_{\text{total}}|a,t) = C_{\text{m}}(a,t) + C_t(a,t) + C_r(a,t) \tag{1}
\]

with:

\[ C_{\text{m}}(\cdot) : \text{function of } a \text{ and } t, \text{ expressing the lifetime-averaged direct costs associated with maintenance action } a \text{ at time } t; \]
\[ C_t(\cdot) : \text{function of } a \text{ and } t, \text{ expressing the lifetime-averaged indirect costs of maintenance (e.g., related to downtime) associated with action } a \text{ at time } t; \]
\[ C_r(\cdot) : \text{function of } a \text{ and } t \text{ expressing the costs associated with the risk of action } a \text{ being inadequate or time } t \text{ being too late.} \]

The risk costs \( C_r(a,t) \) can be further specified as:

\[
C_r(a,t) = \sum_{i=1}^{p} \sum_{j=1}^{\ell} P(H(t) = f_j) P(F_i(t) = 1|H(t) = f_j) C_{ci} + \sum_{j=1}^{\ell} P(H(t) = f_j) C_{fi}(a) \tag{2}
\]

with:

\[ C_{ci} : \text{additional costs of a failure in mode } f_i; \]
\[ C_{fi}(\cdot) : \text{function of } a \text{ expressing the penalty costs of preparing a (wrong) maintenance activity } a \text{ in the case of fault type } f_i. \]

The first term expresses the costs related to the risk of maintenance time \( t \) being too late to avoid a particular failure, with \( F_i(t) \) a binary variable indicating whether the system fails in mode \( i \) at maintenance time \( t \). The second term expresses the costs related to the risk of maintenance action \( a \) being not appropriate to repair the system.
Based on the expected costs, the expected utilities can be defined as:

\[
\mathbb{E}(u(a, t)) = -\mathbb{E}(C_{\text{total}}(a, t))
\]

\[
\mathbb{E}(u(a, t)) = -\left( \sum_{i=1}^{p} \sum_{j=1}^{t} P(H(t) = f_j) P(F_i(t) = 1|H(t) = f_j)C_{e_i} + \sum_{j=1}^{t} P(H(t) = f_j)C_{f_j}(a) + C_{m}(a, t) + C_{i}(a, t) \right)
\]

To compute \(\mathbb{E}(u(a, t))\), next to the cost functions, the probabilities \(P(H(t) = f_j)\) and \(P(F_i(t) = 1|H(t) = f_j)\) need to be known for \(j = 1, \ldots, \ell\) and \(i = 1, \ldots, p\). The probability that a certain fault is present, i.e. \(P(H(t) = f_j)\), reflects the diagnostic result. The probability that the system fails at a particular maintenance time given the system health state, i.e. \(P(F_i(t) = 1|H(t) = f_j)\), refers to prognostic information.

2) Specification of the prognostic result: From the analysis of the maintenance optimization process, we conclude that, in the case of multiple degradation and failure modes, the prognosis process should output conditional predictions of the system reliability, i.e. the functions \(P^j_{\text{fail},i}(\tau)\) defined by:

\[
P^j_{\text{fail},i}(\tau) = P(F_i(\tau) = 1|H(\tau) = f_j),
\]

where \(P^j_{\text{fail},i}(\tau)\) indicates the probability of a failure in mode \(c_i\) at time \(\tau\) given that the system degrades according to mode \(d_j\).

Remark 1: In agreement with the above result, we propose the set of conditional system reliabilities as prognostic measure. Considering conditional system reliabilities allows to account for different costs associated with different failure modes and maintenance activities in the subsequent decision making process. Consequently, in the case of multiple degradation and failure modes, this measure is more valuable than one overall estimation of the remaining useful life.

VI. DEGRADATION MODELING AND FORECASTING

As a first step to determine the conditional predictions of the (future) system reliability (see Section V), in this section, we propose a stochastic filtering approach for degradation modeling and forecasting. Later, in Section VII, we use these predictions to determine (future) system reliability.

Note that the strategies presented in the rest of this paper paper are also valid when another forecasting model (e.g. one based on gamma processes [14], [15]) is used as long as the model outputs a distribution of the degradation measure (and not just the expected value).

A. Multivariate, multiple-model degradation modeling

For each fault \(f_j\), \(j = 1, \ldots, \ell\), the corresponding time behavior of a \(\ell\)-dimensional degradation process \(\{X(\tau) = [X_1(\tau), \ldots, X_\ell(\tau)]^T, \tau \geq 0\}\) is described by a Wiener process\(^3\) plus (nonlinear) drift, i.e.

\[
X(\tau) = m_j(\tau, \theta_j(\tau)) + \sigma_j B(\tau)
\]

with \(\sigma_j B(\tau) = [\sigma_{j,1}, \ldots, \sigma_{j,\ell}]^T B(\tau)\) a Wiener process, i.e. \(B(\tau)\) represents a standard Brownian motion with \(\sigma_j B(\tau) \sim N(0, \sigma_j^2 \tau)\). Models \(m_j\) till \(m_\ell\) are \(\ell\)-dimensional vectors the elements of which are (nonlinear) mappings expressing non-decreasing degradation trends (e.g. linear [9], exponential [1], quasi-linear/asymptotic [29]) associated with the corresponding fault mode \(f_j\). The vector \(\theta_j(\tau) \in \mathbb{R}^{\ell\times\ell}\) defines the model parameters, which might be stochastic. Here we assume \(\theta_j(\tau) \sim N(\mu\theta_j, \Sigma\theta_j)\).

Information regarding the degradation process is obtained through noise-disturbed measurements, i.e.:

\[
Y(\tau) = X(\tau) + \epsilon(\tau)
\]

with \(\{Y(\tau) = [Y_1(\tau), \ldots, Y_\ell(\tau)]^T, \tau \geq 0\}\) the process describing the time behavior of the measurements, and \(\epsilon(\tau) = [\epsilon_1(\tau), \ldots, \epsilon_\ell(\tau)]^T\), with \(\epsilon_j(\tau) \sim N(0, \gamma_j^2)\). It is assumed that the random variables \(\epsilon, \theta_j, \text{ and } B(\tau)\) are mutually statistically independent.

The proposed degradation model (5)-(6) can describe a wide range of degradation trends, and captures both temporal, sampling, and measurement uncertainty [9]. Temporal uncertainty, which is the uncertainty associated with the progression of the degradation over time, is characterized by the dynamics of the Brownian motion \(B(\tau), \tau \geq 0\). Sampling (or case-to-case) variability characterizes the heterogeneity among the degradation paths of different systems under different operation conditions, and is represented in (5) by the stochastic parameters \(\theta_j(\tau)\). Finally measurement uncertainty is reflected by the error term \(\epsilon(\tau)\) in (6), and reflects the fact that the degradation cannot be perfectly measured, i.e. the measurements are disturbed by measurement errors arising e.g. from non-ideal measurement instruments. Moreover, in our modeling framework, modeling uncertainty is minimized by considering a separate model for each fault cause.

B. Online updating and forecasting

Suppose the degradation process is monitored at times \(\tau_1 < \tau_2 < \tau_3 < \ldots\) and let \(Y_k = Y(\tau_k)\) denote the observation vector at time \(\tau_k\). The sequence of measurement vectors \(Y_1, Y_2, \ldots, Y_k\) is represented by \(Y_{1:k}\) and the corresponding sequence of degradation measures is represented by \(X_{1:k}\), with \(X_k = X(\tau_k)\). At time \(\tau_k\) the goal is to estimate the current degradation state \(X(\tau_k)\) and to predict the evolution of the degradation measure \(X(\tau_q)\) for \(\tau_q > \tau_k\) based on the model (5)-(6) and observations \(Y_{1:k}\). For that purpose, we rewrite the model (5)-(6) as a discrete-time stochastic state space model:

\(^3\)Wiener processes are considered because they can model non-monotonic degradation behavior, which is often encountered in practice [1], [9], [36]. In case of monotonic degradation behavior, gamma or compound Poisson processes can be used instead.
Fig. 4. Bayesian filtering. At each time step $k$, first, the state is estimated based on the model (prediction step). Next, this estimate is updated based on the current measurements $Y_k$ (correction step).

\[
\begin{align*}
X_{j,k}^i &= \left( X_{j-1}^i + m_j(\tau_k, \theta_{j,k-1}) - m_j(\tau_{k-1}, \theta_{j,k-1}) + v_{j,k}^i \right) \\
\theta_{j,k}^i &= \left( \theta_{j,k-1} + v_{j,k}^i \right)
\end{align*}
\]

\[
Y_k = X_{j,k} + \epsilon_k, \quad \text{for system in degradation mode } j
\] (7a)

\[
Y_{j,k} = C X_{j,k-1}^i + \eta_{j,k}^i
\] (9a)

with $S_{j,k}^i \in \mathbb{R}^{z_n \times n_m}$, the state vector at $\tau_k$ according to model $j$, which is composed of the degradation measure $X_{j,k}^i \in \mathbb{R}^z$ and the parameter vector $\theta_{j,k}^i \in \mathbb{R}^{n_m}$, $Y_k \in \mathbb{R}^z$ the measurements, $v_{j,k}^i \sim N(0, \text{diag}(\sigma_{j,k}^2_1(\tau_k - \tau_{k-1}), \ldots, \sigma_{j,k}^2_{n_f}(\tau_k - \tau_{k-1})))$, and $\epsilon_k$ the realization of $\epsilon$ at $\tau_k$. Equation (7a) is the transition equation, which specifies how each element of the state vector evolves over time according to degradation model $j$. Equation (7b) is the output equation, specifying how the measurements are linked to the system states. In this formulation, degradation forecasting can be considered as a state estimation and prediction problem, where the goal is to estimate and predict the state, so to statistically minimize the state error. This is a common problem that can be solved using Bayesian filtering [37]. The Bayesian approach to statistics attempts to utilize all available information, i.e. it combines new information with existing knowledge, in order to reduce uncertainty. The formal mechanism to combine new information with existing knowledge is known as Bayes’ theorem [37]. Roughly, this information fusion consists of two steps: a prediction step based on the state transition equation, and a correction step based on new measurements (see Figure 4).

Different types of Bayesian filters have been proposed, among which the Kalman filter [38], its nonlinear extensions, i.e. the extended and unscented Kalman filter, and particle filters [39]. The choice for a filter depends on the exact form of the model (7) and on computational constraints. When the transition and output equation are linear in the state, and the process and measurement noise are additive and Gaussian, the Kalman filter is the optimal filter. When the linearity assumptions are violated (but the noise assumptions are satisfied), an extended or unscented Kalman filter can be used. Another possibility is to use a particle filter, a Monte Carlo methodology, which can also be used when the noise is non-Gaussian or non-additive. The performance of a particle filter depends on the number of particles used.

In general, when enough particles are used, a particle filter outperforms the extended and unscented Kalman filter in terms of estimation accuracy and robustness, but at the costs of higher computational demands [40]–[42].

Because of the attractive properties of the Kalman filter (e.g. computational efficiency, analytic solutions), work has been devoted to transform nonlinear degradation data to an approximate linear form. Examples of such transformations are the log transformation [1] and the time-scale transformation [43]. This way, analytic solutions can be obtained for the approximate linear degradation process in a computationally efficient way, however, at the cost of modeling error. So, for nonlinear degradation processes a trade-off needs to be made between modeling accuracy and solution accuracy. This trade-off is application-specific and a further elaboration is beyond the scope of this paper.

For clarity of presentation, in the remainder, we consider the case that the degradation process can be accurately described by a linear stochastic state space model as considered in [21], i.e. for all $j$, model $m_j$ is of the form:

\[
m_j(\tau_k, \theta_{j,k}) = \beta_j(\tau_k, \phi_j) \theta_{j,k}
\] (8)

with $\phi_j \in \mathbb{R}^{n_{\phi_j}}$, a vector of deterministic parameters, $\theta_{j,k} \in \mathbb{R}^{n_{\theta_j}} \sim N(\mu_{\theta_j}, \Sigma_{\theta_j})$ a vector of stochastic parameters, and $\beta_j$ a $z \times n_{\gamma_j}$ matrix. In this case, model (7) can be written in a linear form:

\[
S_{j,k}^i = \left( X_{j,k}^i \right) = A_{j,k} S_{j,k-1}^i + \eta_{j,k}^i
\] (9a)

\[
Y_{j,k} = C S_{j,k}^i + \epsilon_k
\] (9b)

with:

\[
A_{j,k} = \begin{bmatrix} I & \beta_j(\tau_k, \phi_j) - \beta_j(\tau_{k-1}, \phi_j) \\ 0 & I \end{bmatrix}
\]

\[
\eta_{j,k}^i = \begin{bmatrix} v_{j,k}^i \\ 0 \end{bmatrix} \sim N(0, Q_{j,k})
\]

\[
Q_{j,k} = \begin{bmatrix} \text{diag}(\sigma_{j,k}^2_1(\tau_k - \tau_{k-1}), \ldots, \sigma_{j,k}^2_{n_f}(\tau_k - \tau_{k-1})) & 0 \\ 0 & 0 \end{bmatrix}
\]

\[
C = \begin{bmatrix} I & 0 \end{bmatrix}
\]

\[
\epsilon_k \sim N(0, R)
\]

\[
R = \text{diag}(\gamma_1^2, \ldots, \gamma_z^2)
\]

Procedure 1 outlines the degradation estimation and forecasting based on the Kalman filter. Note that at this stage, we just predict the values of all degradation measures according to all degradation modes. When computing the system reliability (see Section VII), the information from the different degradation measures will be combined. The information from the different fault-specific models will be merged only in the maintenance optimization (see Section VIII).
**Procedure 1** Multiple-model degradation estimation and prediction at time $\tau_k$.

**Input:** Previous states $S^j(k|k-1)$, previous covariance matrices $P^j(k|k-1)$, and matrices $A_{j,k}$ and $Q_{j,k}$, for $j = 1, \ldots, \ell$; matrices $C$ and $R$; failure criteria.

1: Measure $Y_k$
2: for $j = 1, \ldots, \ell$ do

**Estimation of current degradation**

3: Prediction step:
\[
S^j(k|k-1) = A_{j,k}S^j(k-1|k-1) \\
P^j(k|k-1) = A_{j,k}P^j(k-1|k-1)A_{j,k}^T + Q_{j,k}
\]

4: Correction step:
\[
K^j(k) = P^j(k|k-1)C^T(CP^j(k|k-1)C^T + R)^{-1} \\
S^j(k) = S^j(k|k-1) + K^j(k)(Y_k - CS^j(k|k-1)) \\
P^j(k) = \left( I - K^j(k)C \right)P^j(k|k-1)
\]

**Prediction of future degradation**

5: $n \leftarrow 0$
6: while failure criteria are not met do
7: $n \leftarrow n + 1$
8: $n$-step ahead prediction:
\[
S^j(k+n|k) = (A_{j,k})^n S^j(k|k) \\
P^j(k+n|k) = (A_{j,k})^n P(k|k)(A_{j,k}^T)^n + \sum_{l=0}^{n-1} (A_{j,k})^lQ_{j,k}(A_{j,k}^T)^{n-l}
\]
9: end while
10: end for

**Output:** predictions of the degradation measure $X^j(\tau_q)$ according to all degradation modes $d_j$ for $q = k, k+1, \ldots, k+n$

**VII. System reliability**

**A. Multivariate definitions**

Two prognostic measures are (future) system reliability and remaining useful life [4], [9], [25], [44]. Although the remaining useful life is most commonly used, in this paper we focus on the system reliability. We made this choice because this measure best fits the subsequent maintenance optimization process (see Section V). Before elaborating on the system reliability, we define system failure in the multivariate case.

In the univariate case, system failure is usually defined as:

$$X(\tau) \begin{cases} < \lambda \Rightarrow \text{system is functional at } \tau \\ \geq \lambda \Rightarrow \text{system fails at } \tau \end{cases}$$

(10)

Failure definition (10) can be extended to the multivariate case by defining a failure in mode $c_i$ as $g_i(X(\tau))$ being larger than or equal to a predefined threshold $\lambda_i$, i.e.:

$$g_i(X(\tau)) \begin{cases} < \lambda_i \Rightarrow \text{no failure in mode } c_i \text{ at } \tau \\ \geq \lambda_i \Rightarrow \text{system fails in mode } c_i \text{ at } \tau \end{cases}$$

(11)

with $g_i(\cdot)$, $i = 1, \ldots, p$, application-specific functions, which we will elaborate on in Section VII-B. Accordingly, system failure is defined as:

$$g_i(X(\tau)) < \lambda_1 \text{ and... and } g_p(X(\tau)) < \lambda_p \Rightarrow \text{functional at } \tau \text{ and... and...} \text{ or } g_p(X(\tau)) \geq \lambda_p \Rightarrow \text{failure at } \tau$$

(12)

The system reliability ($P_{\text{func}}$) is defined as the probability that the system is functional, i.e. does not fail [25]. Based on (11), the system reliability with respect to failure mode $c_i$ at time $\tau$ is the probability that $g_i(X(\tau))$ is smaller than $\lambda_i$, i.e.

$$P_{\text{func},i}(\tau) = p\left(g_i(X(\tau)) < \lambda_i \right)$$

(13)

The overall system reliability at time $\tau$ is defined as the probability that the system is functional, i.e. is not in any of the failure modes $c_1 \text{ till } c_p$:

$$P_{\text{func}}(\tau) = P\left(g_1(X(\tau)) < \lambda_1, \ldots, g_p(X(\tau)) < \lambda_p \right)$$

(14)

From these definitions, we conclude that the functions $P_{\text{fail},i}$ as defined in (4) are related to predictions of the system reliability with respect to failure mode $i$ conditional to degradation mode $j$. So, in accordance with the above definitions, (4) can be written as:

$$P_{\text{fail},i}(\tau) = 1 - P_{\text{func},i}(\tau)$$

(15)

with

$$P_{\text{func},i}(\tau) = p\left(g_i(X(\tau)) < \lambda_i \right)$$

**B. Determination of failure definition and system reliability**

The functions $g_i(\cdot)$, $i = 1, \ldots, p$, used to define system failure (11) are application-specific and depend on the relationships between the degradation measures $X_\xi$, $\xi = 1, \ldots, z$. Here, we focus on three common types of relationships (see Figure 5 for 2-D example relationships):

1) complementary measures:
   a) independently assessable;
   b) not independently assessable;

2) redundant measures.

Measures $X_\xi$, $\xi = 1, \ldots, z$, are complementary and independently assessable if it holds that the system is functional in mode $c_i$ if and only if each measure $X_\xi$ is below an individual threshold $\lambda_i$. In the case of redundant measures, only $k$ out of $z$, $k < z$, of components $X_\xi$ need to be below their threshold $\lambda_i$ for the system to be functioning in mode $c_i$. For complementary, not independently-assessable measures, no relevant individual thresholds exist. For the system to be functioning, all functions $g_i(\cdot)$ of $X$ should then just be below an overall threshold value $\lambda_i$. 
Fig. 5. 2-D illustration of three types of failure definitions. (a) independently-assessable complementary measures, (b) redundant measures, (c) not independently-assessable complementary measures.

For brevity, in the sequel we omit the subscript \( i \) whenever the explicit reference to a particular failure mode is not necessary. For the same reason, we omit the time argument \( \tau \) whenever possible.

1) Independently-assessable complementary measures: For measures that are complementary and independently assessable, the functions \( g_i(\cdot), i = 1, \ldots, p \), can be chosen arbitrarily, as long as they satisfy:

\[
g_i(X) \begin{cases} 
\geq \lambda_i & \text{if } \max(X_1 - \lambda_{i,1}, \ldots, X_z - \lambda_{i,z}) \geq 0 \\
< \lambda_i & \text{otherwise}
\end{cases}
\]  

The system reliability with respect to failure mode \( c_i \) is computed as:

\[
P_{\text{unc}, i} = \int_{-\infty}^{x_{1,1}} \int_{-\infty}^{x_{1,2}} \cdots \int_{-\infty}^{x_{1,z}} p(X_1, X_2, \ldots, X_z) dX_1 dX_2 \cdots dX_z
\]  

with \( p(\cdot) \) the distribution function of the degradation measure \( X \), which follows from the degradation modeling and forecasting (see Section VI).

2) Redundant measures: Safety-critical systems are often equipped with redundancy, e.g. airplanes having more engines than necessarily for take-off. Systems can be redundant to vary degrees. The redundancy is lowest when \( z-1 \) out of \( z \) components need to be functioning for the whole system to be functioning, and highest when only \( 1 \) out of \( z \) components needs to be functioning for the whole system to be functioning. If the functioning of each redundant component is reflected by one degradation measure \( X_\xi \), then for a \( k \)-out-of-\( z \): \( G \) system\(^4\) at least \( k \) out of \( z \) measures \( X_\xi, \xi = 1, \ldots, z \), need to be below their threshold \( \lambda_{i,\xi} \) for the system to be functioning in mode \( c_i \). So, \( g_i(\cdot) \) should be chosen such that:

\[
g_i^{(k)}(X) \begin{cases} 
< \lambda_i & \text{if } \sum_{\xi=1}^{z} \alpha_{i,\xi} < k \\
\geq \lambda_i & \text{otherwise}
\end{cases}
\]  

with:

\[
\alpha_{i,\xi} = \begin{cases} 
1 & \text{if } X_\xi < \lambda_{i,\xi} \\
0 & \text{otherwise}
\end{cases}
\]

\(^4\)A \( k \)-out-of-\( z \): \( G \) system is a system that works well if at least \( k \)-out-of-\( z \) components work well.

and the superscript \( (k) \) indicating that we consider a \( k \)-out-of-\( z \): \( G \) system. The system reliability with respect to failure mode \( c_i \) is computed as:

\[
P_{\text{unc}, i}^{(k)} = \int \cdots \int p(X_1, X_2, \ldots, X_z) dX_1 dX_2 \cdots dX_z
\]

with \( \Omega_{i}^{(k)} = \Omega_{i,1}^{(k)} \cup \cdots \cup \Omega_{i,p}^{(k)} \) the integration surface representing the subset of \( X \in \mathbb{R}^z \) for which the system is not in failure mode \( c_i \) and \( \tau = 1, \ldots, k \) the different configurations for which \( k \) degradation measures are in their desired region, with \( \Omega_{i,\tau}^{(k)} \) the corresponding surfaces. The last term \( kR_i^{(k+1)} \) corrects for the overlap between the integration surfaces associated with the different configurations \( \tau = 1, \ldots, k \). To illustrate, Figure 6 gives the integration surfaces \( \Omega_{i}^{(3)}, \Omega_{i}^{(2)}, \) and \( \Omega_{i}^{(1)} \) for a three-dimensional case, which are defined as:

\[
\Omega_{i}^{(3)} = \{X \in \mathbb{R}^3 : X_1 < \lambda_1 \text{ and } X_2 < \lambda_2 \text{ and } X_3 < \lambda_3\}
\]

\[
\Omega_{i}^{(2)} = \{X \in \mathbb{R}^3 : (X_1 < \lambda_1 \text{ and } X_2 < \lambda_2) \text{ or } (X_1 < \lambda_1 \text{ and } X_3 < \lambda_3) \text{ or } (X_2 < \lambda_2 \text{ and } X_3 < \lambda_3)\}
\]

\[
\Omega_{i}^{(1)} = \{X \in \mathbb{R}^3 : (X_1 < \lambda_1 \text{ or } X_2 < \lambda_2 \text{ or } X_3 < \lambda_3)\}
\]

3) Not individually-assessable complementary measures: In practice, it is common that the functioning of a system is defined as a combination of the degradation measures satisfying a certain criterion, e.g. the sum or product of measures \( X_\xi, \xi = 1, \ldots, z \), should be below a threshold. In such situations, system reliability cannot be assessed based on individual threshold values. However, in such cases, the critical surface is generally smooth and can be written in the form:

\[
s_{\text{cr}}(X_1, X_2, \ldots, X_z) = 0
\]  

with \( s_{\text{cr}}(\cdot) \) a continuous function (see Figure 7 for some two-dimensional example surfaces and the associated functions.
Fig. 6. 3-D visualization of the integration surfaces indicating the subsets of $X \in \mathbb{R}^3$ for which the system is functional: (a) 3-out-of-3: G system; (b) 2-out-of-3: G system, (c) 1-out-of-3: G system.

\[(X_1 - c_1)^2 + (X_2 - c_2)^2 - c_3^2 < 0\]

Fig. 7. 2-D example surfaces indicating the subset of instances of $X = [X_1 X_2]^T$ for which the system is functional.

\[s_{cr}(\cdot)\). In this case, the integration bounds directly follow from $s_{cr}(\cdot)$.

4) Concluding remark: In the multivariate case, the failure definition and the associated computation of the system reliability depend on the relationships among the degradation measures. Three common relationships have been discussed. In general, the dependencies among degradation measures do not always fall within one category. Consider for example a system with four degradation measures $X_1$ till $X_4$ and failure defined as:

\[g(X_1, X_2) > \lambda_1 \text{ and } g'(X_3, X_4) > \lambda_2 \implies \text{system failure} \] (21)

For this system, we have to deal with both redundant and not individually-assessable complementary measures. In such cases the strategies discussed before can be combined.

VIII. EVALUATION AND DISCUSSION

A realistic and thorough evaluation of the proposed approach is only possible for a particular application and in combination with a fault diagnosis and maintenance optimization approach. Such an evaluation goes beyond the scope of this paper. In this section, we will reflect on two main attributes of the proposed approach, namely:

1) its position within the condition-based maintenance process;
2) the added value of using multiple models over using a single model on the prediction quality, and its dependence on the diagnostic result.

A. Position within the condition-based maintenance process

Procedure 2 summarizes the proposed prognosis strategy within a condition-based maintenance process. Although we ensure that the different processes are compatible with each other in the sense that the diagnostic and prognostic results support maintenance optimization, we do not impose further requirements on the diagnosis and maintenance optimization process. In particular, we only assume that the diagnosis process outputs a probability distribution over the system health state, and that decision making is done based on expected utilities. Even when the diagnostic result is specified using another uncertainty formalism (e.g. possibilities, fuzzy measures, mass functions) the proposed strategy is of use. In this case the alternative uncertainty distribution first has to be transformed into a probability distribution. For this task, transformation rules are available in literature [45], [46]. Moreover, if desired, another multivariate multiple-model forecasting algorithm can be used instead of the forecasting strategy outlined.
in Procedure 1. For example a multiple-model method based on gamma processes [14], [15] in the case that degradation behavior is best described by a monotonic process. We regard the freedom to independently select an optimal diagnosis and forecasting algorithm as a practical advantage. Indeed, problem characteristics vary widely among applications, and so the best combination of diagnosis and prognosis approach is highly application-specific. As another advantage, we regard the fact that the maintenance optimization model we rely on is based on cost functions that are easily assessed by practitioners (e.g., costs of maintenance, costs associated with a failure, costs associated with downtime).

Procedure 2 Prognosis within condition-based maintenance at time \( \tau_k \).

**Input:** Failure functions \( g_i() \) and thresholds \( \lambda_i \) for \( i = 1, \ldots, p \), set \( T \) of possible maintenance time instants

**Fault diagnosis**

\begin{align*}
1: \text{generates } P(H(\tau_k))
\end{align*}

**Prognosis**

\begin{align*}
2: \text{for } t \in T, t \geq \tau_k \text{ do}
3: \quad \text{for } j = 1, \ldots, t \text{ do}
4: \quad \quad \text{Determine } X^j(t) \text{ using Procedure 1}
5: \quad \text{for } i = 1, \ldots, p \text{ do}
6: \quad \quad \text{Determine } P^j_{i, \text{func}, i}(t):
7: \quad \quad \quad P^j_{i, \text{func}, i}(t) = \int \int \cdots \int \Omega_i \in X \text{ the surface for which } g_i(X) < \lambda_i
8: \quad \text{end for}
9: \quad \text{end for}
10: \text{Based on } P(H(\tau_k)) \text{ and } P^j_{i, \text{func}, i}(t) \forall t \in T, t \geq \tau_k, \forall j \in \{1, \ldots, t\}, \forall i \in \{1, \ldots, p\}, \text{determine the optimal maintenance decision, e.g. according to (3)-(4)}

**Output:** Maintenance decision for \( \tau_k \)

**B. Multiple models**

We motivated our choice for a multiple-model approach by the fact that system degradation may evolve differently over time for different types of faults. For example, for the railway case (see Figure 3) the ability to detect a free track decreases approximately linearly over time in the case of an electrical insulation problem, while the temporal degradation behavior is best described by an exponential model when a rail surface defect is present. Therefore, a natural choice is to use a linear model to forecast degradation resulting from an insulation problem, while using an exponential model to forecast degradation behavior as a consequence of a rail surface defect.

We conclude that a multiple-model approach has the potential to outperform a single-model approach with respect to prediction accuracy. We say “has the potential to” because the actual prediction performance of a multiple-model approach heavily relies on knowledge of the underlying degradation mode. In practice, we do not know with certainty which fault is present, and so which model best describes degradation behavior. This means that for online degradation forecasting the potential improvement in prediction accuracy cannot be fully utilized. Whether and to what extent a multiple-model approach will outperform a single-model approach with respect to prediction accuracy depends, among other things, on the accuracy of the diagnostic result. Although promising fault diagnosis methods have been proposed in the literature (see e.g. [47]–[49]), the achievable diagnostic accuracy is rather application-specific. Moreover, in general, diagnostic quality improves with the severity of the fault; so for incipient faults, diagnostic quality may be low.

Figure 8 shows two typical temporal behaviors of the diagnostic result for gradually evolving faults taken from [50]. These behaviors relate to a track circuit diagnosis task for which a Kalman filter approach has been used. In both Figures 8(a) and (b), the system is healthy till \( \tau_d \), i.e. \( H(\tau) = h \) for \( \tau < \tau_d \). Afterwards, the system suffers from fault \( f_2 \), i.e. \( H(\tau) = f_2 \) for \( \tau \geq \tau_d \). From the diagnostic results in Figure 8, we conclude that in both cases the presence of a fault is almost instantly detected, i.e. \( P(H(\tau) = h) \approx 0 \) for \( \tau > \tau_d \). However, only from \( \tau = \tau_i \) on the system behavior is adequately diagnosed. In Figure 8(a), the fault is initially misdiagnosed, i.e. we conclude with a probability of approximately 80% that the system suffers from fault \( f_1 \). Slightly later, when more data have been collected, fault \( f_2 \) is correctly identified. In Figure 8(b) for \( \tau \) between \( \tau_d \) and \( \tau_i \) there is (much) uncertainty about the cause of the faulty behavior. Initially all faults are plausible, i.e. both \( P(H(\tau) = f_1) \), \( P(H(\tau) = f_2) \), and \( P(H(\tau) = f_3) \) are significantly larger than zero. Afterwards, doubt remains between faults \( f_2 \) and \( f_3 \) only. From time \( \tau_i \), the actual fault cause is identified with high accuracy, i.e. \( P(H(\tau) = f_2) \approx 1 \).

In general, the longer the fault is present, the more data of the faulty behavior are available, and the more accurate and reliable the diagnostic result is. How long it takes before adequate diagnostic results are obtained is however rather application-specific. Since in general both diagnostic and degradation forecasting performance improve over time, it is important to account for this time behavior in the subsequent decision making process.

**IX. Conclusions**

We have proposed a multiple-model approach to degradation modeling and forecasting for systems with multiple degradation and failure modes. A stochastic filtering approach is considered to handle the different sources of uncertainty inherent to degradation forecasting. Moreover, the links with the other tasks of the condition-based maintenance process, i.e. diagnosis and maintenance planning, have been established. We conclude that conditional predictions of future system reliability best support the subsequent decision making process. Accordingly, a framework has been proposed to determine the (future) system reliability in the multivariate case for different types of relationships among degradation measures.
We conclude that by using multiple models to forecast degradation behavior, the modeling error can be reduced. However, since the applicable model is selected based on the diagnostic result, the benefit of using multiple models over using a single model highly depends on the accuracy of the diagnostic result. Given the current quality of diagnosis methods, we do not expect this to be a serious drawback. However, caution is needed when faults are in their incipient phase. In this phase, the diagnostic results are often less accurate. A thorough analysis of the accuracy of diagnosis and prognosis results over time, and its implications on the subsequent maintenance optimization process is therefore an interesting topic for further research. As another topic for further research, we propose the thorough evaluation of the proposed approach within a condition-based maintenance process.

ACKNOWLEDGMENT
This research is part of the STW/ProRail project “Advanced monitoring of intelligent rail infrastructure (AD-MIRE)”, project 12236, which is supported by ProRail and the Dutch Technology Foundation STW, which is part of the Netherlands Organization for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no. 324432 (AMBI project).

REFERENCES


**APPENDIX**

**Expected-utility theory** [34] provides a framework for determining the optimal action given probabilistic information regarding the situation you are in. Its two main ingredients are:

1) Utilities, which indicate the desirability of a particular action in a particular situation, i.e. utilities express preferences among the available choices.

2) Probabilities, which indicate how likely a particular situation is.

The expected utility $E(u|d)$ of a decision $d \in \Theta_D$ is computed as:

$$ E(u|d) = \sum_{v \in \Theta_V} P(v) u(d, v) $$

(22)

with $\Theta_D$ the discrete set of possible decisions, $\Theta_V$ the set of possible situations, $u(d, v)$ the utility of decision $d$ given situation $v$, and $P(v)$ the probability of $v$. Then, an optimal decision $d^*$ is:

$$ d^* = \arg \max_{d \in \Theta_D} E(u|d) $$

(23)