RAMS
Reliability, Availability,
Maintainability, and Safety

HAV6003 - Digital solutions for operation and maintenance of offshore wind farms

Course compendium

Jørn Vatn

April 9, 2024

Department of Mechanical and Industrial Engineering
Norwegian University of Science and Technology

Preface

This course compendium is developed for the course HAV6003 - Digital solutions for operation and maintenance of offshore wind farms. The compendium has been developed with support from Norwegian Directorate for Higher Education and Skills.

The reader is assumed to have background in basic probability theory and some skills in using ICT-tools.

Trondheim, April 9, 2024
Jørn Vatn

Contents

Chapter 1

Introduction

1.1 Background

The digitalization that takes place in the society today will also have an impact within maintenance. Computerized maintenance management systems and advanced signal processing of vibration data have been around for decades. However, more formalized digital models that have predictive capabilities used to support maintenance decision are hardly used in practice. The term 'digital twin' is very popular in these days, and in order to realize the potential of such digital twins within maintenance it is required to establish maintenance models which can be implemented as part of the "twins". A challenge in many companies is the inherent conflict between operations and maintenance. From the operations perspective one would like to produce as much as possible, whereas the maintenance department also require to shut down production in order to carry out maintenance. The idea is that various digital twins can help to sort out some of the conflicting objectives and avoid sub-optimization.

In the literature a huge number of mathematical models exist but they are not often used in practice. Therefore it is important to establish a limited number of mathematical models that covers in a reasonable manner the most common situations.

The following references are based on an NTNU specialization report by Thiruthiyappan (2022): The role of wind turbines (WTs) for producing clean, renewable energy is crucial for reaching the climate goals of the 2015 Paris Agreement. The onshore wind industry is a proven and a mature technology that has an extensive global supply chain and now the offshore wind industry is expected to grow rapidly (Technologies, 2022). The global wind report 2022, published by the Global Wind Energy Council (GWEC) shows that 93.6 gigawatt (GW) of new installations in 2021 brings the global cumulative wind power capacity to 837 GW, showing a year-over-year growth of 12% (Lee and Zhao, 2022). Figure 1.1 depicts the installed capacity of onshore and offshore wind farms from 2017 to 2021 in GW, with a 1.8% decrease from 2020 to 2021 due to the COVID-19 pandemic. In 2020, only 6.9 GW of offshore installations was com-

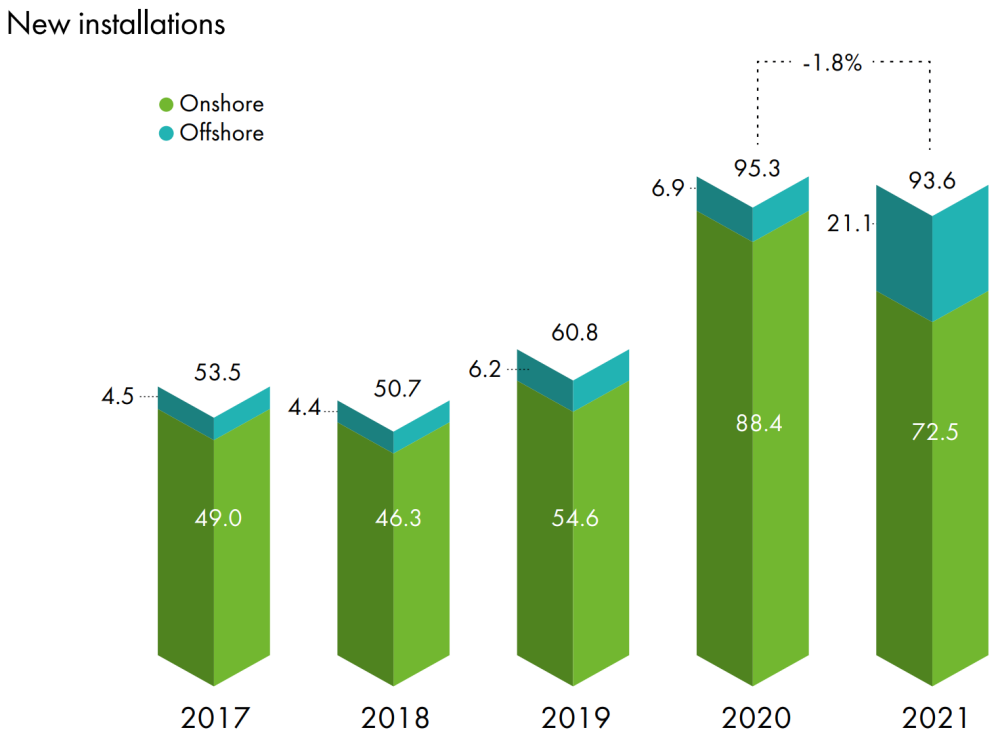


Figure 1.1: New installations onshore and offshore in GW (Lee and Zhao, 2022)

missioned. Whereas, in 2021, 21.1 GW was commissioned, which is a threefold increase in the offshore wind market. The world's total offshore capacity reached 57 GW, which is 7% of all global installations (Lee and Zhao, 2022). This suggests a strong trend and higher offshore wind farm capacities to be installed in the upcoming years.

Due to more consistent wind speeds offshore than when compared with onshore locations, deploying larger WTs with higher capacity in the sea can take advantage of this (Technologies, 2022). Hence, greater wind power generation is achievable. However, the harsh weather conditions, largely variable aerodynamic, gravitational, centrifugal and gyroscopic loads induce higher failures rates, and frequency of faults and failures in the WTs (Badihi et al., 2022). Furthermore, the remote locations of the offshore WTs sites make it difficult challenge to conduct maintenance tasks (Zhang et al., 2022). This is due to greater logistics costs, difficulties with maintenance scheduling in consideration with weather condition uncertainties and or lower skilled manpower (Badihi et al., 2022). Presently, operation and maintenance (O&M) costs account for anywhere from 10% to 30% of the total energy generation cost of onshore WTs, whereas in offshore WTs the O&M costs can surge up to 25% to 50% (Badihi et al., 2022). This issue must be addressed and hence, to upkeep offshore WTs cost-effectively, ensure efficient production and financial viability of wind power, it is crucial to maintain offshore WTs reliability and availability (Badihi et al., 2022). For this predictive maintenance (PdM) is an appealing strategy for

the offshore wind industry (Zhang et al., 2022). PdM aims to monitor the condition of mechanical components and predict the upcoming failures. Implementing the PdM strategy can reduce the unexpected failures of critical offshore WTs components with RUL prediction and increase the production availability of WTs. The number of trips to the sea for maintenance activities can thus be reduced hence improving the safety, the maximum working life of critical WTs components are utilized and catastrophic damages can be avoided (Fox et al., 2022).

1.2 Objectives

The main objectives of this course compendium is to introduce the reader into maintenance theory and apply methods, models and programming techniques to improve maintenance and operations of offshore wind farms. In particular the reader shall:

1. Become familiar with maintenance concepts and mathematical models used in maintenance planning and optimization
2. Understand interaction between maintenance and operations
3. Get familiar with the concept of digital twins and develop simple digital twins demonstrating interaction between maintenance and operations
4. Become motivated to learn more

1.3 ICT-tools and resources

A large number of mathematical models are presented in this compendium. In order to apply these models to support maintenance it is required to have some tools. Most of the models presented can be run from the spreadsheet mode in Excel. To support this, a dedicated Excel fail is presented, i.e., the MaintOp.xlsm file. This file contains build-in code for many of the models presented.

An alternative to run the models from within Excel, it is also possible to implement these models in Python. An introduction to Python is provided for this course, and example files are provided from the Python repository .

A set of solutions to exercises can be found in the Solution repository .

1.4 Outline

Chapter 2 describes the main elements of a computerized maintenance management system (CMMS). A CMMS is a digital tool which is essential for efficient maintenance management.

However, a CMMS is usually a static tool containing information regarding the asset in terms of inventory listing, preventive and corrective maintenance activity conducted, and work orders for future activities to be conducted.

Chapter 3 presents a framework for structuring elements of so-called digital twins. The DNV-RP-A204 (2021) recommended practice is a starting point and then various digital twins are introduced and we discuss how these could interact.

Chapter 4 defines the basic maintenance terminology followed by reliability terminology and concepts. Classical age and block replacement maintenance policies are introduced.

Chapter 5 presents the ideas behind predictive maintenance. We introduce state variables and stochastic processes which are the basis for the maintenance models developed. The starting point is the classical PF-model, but also the Wiener and gamma processes are introduced. The concept of remaining useful lifetime (RUL) is defined.

Chapter 6 introduces a model where degradation could be defined in terms of a finite number of degradation levels. Markov theory is applied in order to establish the required maintenance models.

Usually maintenance activities are grouped together to save the so-called setup cost and Chapter 8 presents a framework for grouping of maintenance activities and how to include opportunistic maintenance in the optimization process.

The criticality of failures depends significantly on the downtime after a failure. Downtime again depends on the spare part strategy, and Chapter 9 introduces several models for spare part management.

For prioritization of renewal projects and modifications we have a long-term perspective of investments and net present values (NPV) is often used as a way to distinguish between future costs and cost we pay now. Chapter 10 introduce basic concepts and formulas used in life cycle cost (LCC) analysis.

All models presented in this course require numerical values for the model parameters. Chapter 11 gives an introduction to reliability data analysis. The starting point is life time data analysis, but also methods for estimating model parameters in the Markov model is presented.

In recent years artificial intelligence (AI) and machine learning (ML) have been introduced into more and more application area. Chapter 12 introduces some basic ML-concepts. AI is not covered in this presentation.

The main focus in this course compendium is on maintenance modelling and mathematical modelling to support maintenance optimization. However, to establish a complete maintenance plan we can not afford to establish a detailed maintenance model for all items. Reliability centred maintenance (RCM) is a holistic approach to establish a preventive maintenance program and Chapter 13 presents the main ideas of RCM.

Chapter 2

Computerized maintenance mangement system

2.1 Introduction

A computerized maintenance management system (CMMS) is software that centralizes maintenance information and facilitates the processes of maintenance operations. The folloing section is mainly based on Pedersen (2020). To understand the objective of a CMMS it is fundamental to understand what is maintenance management. Maintenance management is *all activities of the management that determine the maintenance objectives, strategies and responsibilities, and implementation of them by such means as maintenance planning, maintenance control, and the improvement of maintenance activities and economics.* In particular:

- Planning, scheduling, and managing maintenance for parts, vehicles, and other essential equipment
- Predicting potential issues and scheduling regular maintenance tasks to eliminate them
- With more real-time data, it is possible to streamline the maintenance process and make it more cost-effective

An important aspect of maintenance management is to understand the maintenance management loop depicted in Figure 2.1:

The core of a CMMS is its database. It has a data model that organizes information about the assets and equipment a maintenance organization has to maintain, as well as materials and other resources to do so. The main objectives of a CMMS are:

- Provide data for decision
- Organize and manage work orders, i.e., task describing maintenance tasks to be executed

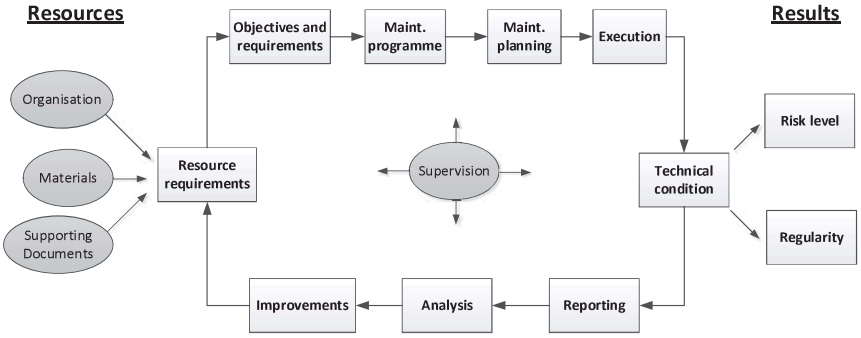


Figure 2.1: Maintenance management loop NORSOK-Z008/HAVTIL

- Report and document the technical condition, and maintenance activities carried out
- Make sure that the maintenance management loop in Figure 2.1 is “closed”

Experience shows that an efficient CMMS enables lower operations and maintenance costs due to

- Efficient information retrieval (overview of assets)
- Efficient platform of communication (Work orders, spares management)

and more explicitly increased availability of a plant or system due to

- Preventive Maintenance just in time
- Lower MLD (efficient information retrieval, efficient communication, efficient spares ordering etc.)
- Lower MRT due to access to documentation
- Lower MDT

where MLD = Mean Logistic Delay, MRT = Mean active Repair Time, and MDT = Mean Down Time.

Important modules of a CMMS are:

- *Equipment register* - for registering company’s equipment / assets
- *Work Order module* - for planning and performing efficient maintenance and modification

- *Preventive Maintenance module* - automatic generation of work orders based on maintenance plan, condition monitoring - notifications
- *Procurement module* - spare parts management,
- *Documentation module* - for handling drawings, procedures, and other information.
- *Analysis module* - for handling statistical data on failures, remaining useful lifetime prediction etc.

2.2 Equipment register

The components are registered with a number in an hierarchic system. The components become and individual object with a history and information linked to it. (NORSOK Z-DP-002). Typical information in the equipment register for an object is:

- Sizes, capacities, operations and maintenance manuals
- Location in the plant
- Preventive maintenance program
- Documentation
- Spares

2.3 Work order module

The work order module is a register of jobs planned and going on in a plant. It is important to have jobs prioritized and done the right way, by skilled personnel using the required methods and procedures.

- Jobs are described in individual work orders with an independent number and, well defined scope, listing of activities with manpower requirements described, listing of spares and tools needed, how to do inspection, what to report, time schedule etc.
- Good job instructions are essential, they shall be detailed enough to have the job done according to requirements. Task analysis if often recommended to develop good job descriptions.
- Include the people going to do the work in writing and outlining the work order if possible.

Requisition control is also an important aspect, i.e., who has the authority to order work and defining the scope. The system will only allow certain people to authorize work and use of resources. A work order shall describe the plan for implementation of the work:

- Time frame for the job
- Personnel amount and type needed
- When must activities happen in time, i.e., a time interval for execution

A work order must also have a closing report, i.e., specification of:

- Time used
- What has been done
- Updated drawings - As Built
- Outstanding work
- Handover reports, signed check lists
- Important findings for continuous improvement

2.4 Preventive Maintenance module

The main purpose of preventive maintenance module is to administrate planned maintenance. The preventive maintenance module shall be able to generate work orders automatically based on maintenance plan, for example established by Reliability Centred Maintenance (RCM). Further alarm lists from condition monitoring system should also be able to generate automatic preventive maintenance work orders. A well defined preventive maintenance work order contains:

- Work description
- Procedures - requirements
- A list of tools needed
- A list of spares
- Plan -time, resources, manpower etc.
- Report requirements, checklists

2.5 Documentation module

The documentation module shall ensure easy access to documentation:

- Beware is the documentation updated and valid ? Can it be trusted?
- Can store photos, videos, paper-based info
- Can extract information online
- Easy retrieval
- Efficient way of communicating

2.6 Analysis module

The main purpose of the analysis module is to support the continuous improvement processes, and reference is made to *Analysis/Improvement* in Figure 2.1. Of particular interest is to review failures, direct failure causes and root causes. It is also important to be able to estimate performance metrics like:

- Failure rates, and the failure rate function
- PF-intervals
- MDT
- MLD
- MRT
- Availability

Chapter 3

Digital twin

3.1 Introduction

A digital twin is a digital representation of a real-world entity or system. The implementation of a digital twin is an encapsulated software object or model that mirrors for example a physical system, historical and future maintenance activities, or an operational plan. Data from multiple digital twins can be aggregated for a composite view. The notion of a digital representation of real-world entities or systems is not new. Its heritage goes back to computer-aided design representations of physical assets or profiles of individual customers. The difference in the latest iteration of digital twins, adopted from Gartner (2019), is:

1. The robustness of the models with a focus on how they support specific business outcomes such that high reliability and efficient maintenance
2. Digital twins' link to the real world, potentially in real-time for monitoring, and control
3. The application of advanced big data analytics and AI/ML to drive new business opportunities
4. The ability to interact with them and evaluate "what-if" scenarios

Experience shows that there is no common definition of the term digital twin, and the aspects to implement in the various companies digital twin varies.

In the recommended practice on qualification and assurance of digital twins (DNV-RP-A204, 2021) it is proposed to define a "ladder" for the evolutional stages of a functional element of a digital twin as depicted in Figure 3.1.

No reported study has been carried out to document evolutionary stages within maintenance. Many companies are implementing digital twins for maintenance, but to our knowledge,

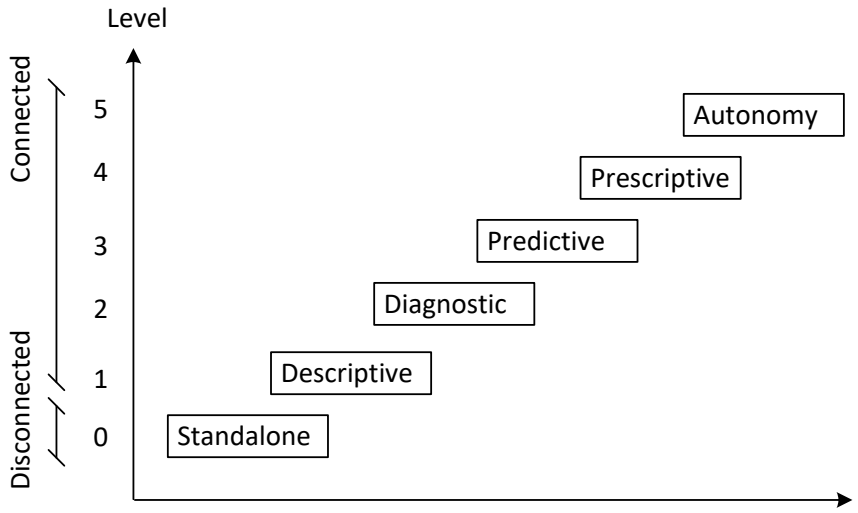


Figure 3.1: Evolution stages or capabilities of a functional element of a digital twin (DNV-RP-A204)

the main effort is on systematizing various systems on the descriptive level and to some extend the diagnostic level.

An objective for offshore wind operation and maintenance is to reach a predictive level for the digital twin implementations according to the stages in Figure 3.1.

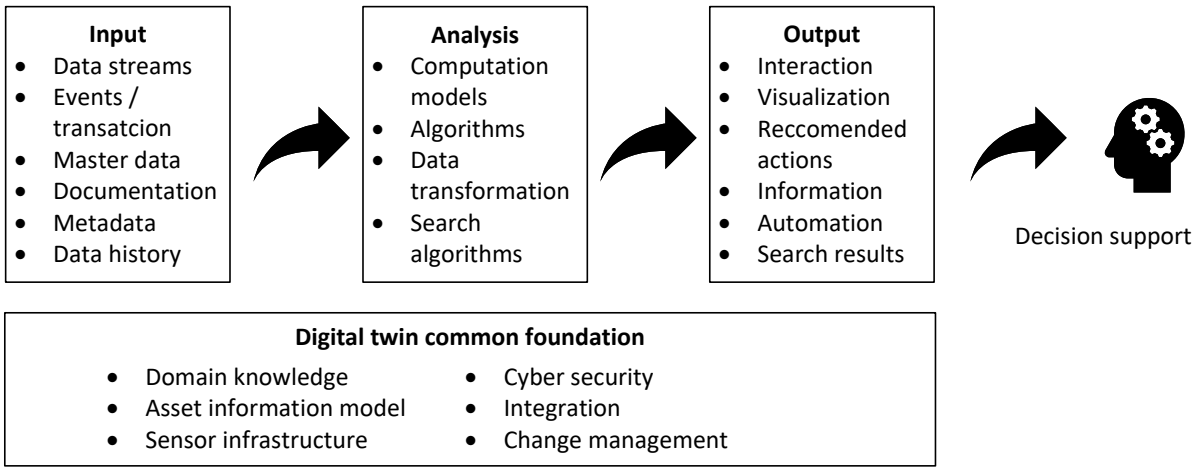


Figure 3.2: Elements of a digital twin (DNV-RP-A204)

DNV-RP-A204 also presents a generic model for the elements of a digital twin. Figure 3.2 indicates a process from input through analysis to output.

To build up the input blocks in Figure 2 we need to integrate several sources, e.g., Buildiing

information management systems (BIM), CMMS, FRACAS and GIS models.

If the ambition also is on the diagnostic and predictive evolution stages the output block needs to be formally linked to the analysis block where input is analysed in (near) real-time.

Finally, the prescriptive and autonomy evolution stages relates to decision support in Figure 3.2, where "support" at the ultimate stage means e.g., automatic generation of maintenance work orders, automatic shutting down of road sections or railway lines in case of bad weather, increased degradation levels etc.

3.2 Digital twin labelling

According to DNV-RP-A204 (2021) a digital twin is defined as a *Virtual representation of a system or asset that calculates system states and makes system information available, through integrated models and data, with the purpose of providing decision support over its lifecycle.* In order to be more explicit regarding functionality of the digital twin we propose to label the digital twins according to the various application domains. The following labelling categories are proposed:

- The *Operations* DT: This DT contains operational plans, cost related to lost production, opportunity windows for maintenance etc
- The *Condition* DT: This DT contains information regarding the condition of the "hardware". In principle this DT contains both current condition, and historical data
- The *Risk* DT: This DT contains the risk picture
- The *Environment* DT: This DT contains information regarding the environment, such as temperature, precipitation, wind etc.
- The *Maintenance cost* DT: This DT contains the relevant mathematical models used for optimizing maintenance, interacting with the Operations DT

We will not elaborate on technical issues, i.e., how these DT are sharing data, how we can conduct "what-if" queries etc.

3.3 Operations DT

The operations DT relates to operational plans, agreed deliverables, logistics etc. For example for a raiolway system the operations DT typically contains the time table, and any planned deviations from the plan. For example freight trains are not necessarily scheduled in advanced. For train operations it is also important to have a model describing how failures and planned maintenance work on the track will affect the throughput, i.e., cancellations, delays etc. For

road segments there will not be any time table, but important information would be daily traffic (ÅDT) split to a required level, i.e., by months and working days vs weekends. For a wind farm the operations DT contains plan for which turbines to operate, blade pitch, yaw drive direction etc.

In order to respond to *short term* what-if queries, it is required to have a model-based foundation and/or data driven based foundation if we have sufficient data for training purposes.

Long term what-if queries would typically be expected changes in traffic volume.

3.4 Condition DT

The condition DT contains information regarding the condition (health) of the assets. Part of such information is contained in a computerized maintenance management system (CMMS). Information from online and offline condition monitoring systems are usually not transferred directly to the CMMS, thus the condition DT also include information from condition monitoring system as well as supervisory control and data acquisition systems (SCADA). Thus, the basic information contained in a DT is:

- The plant hierarchy with the relevant objects
- Information regarding events for each object, in particular preventive maintenance and corrective maintenance tasks
- Condition information from condition monitoring systems, both off-line and on-line systems
- Information from the SCADA system

In this context we include both the information about the asset, e.g., a plant hierarch with the relevant objects (physical items) and the information from condition monitoring systems, and any event information related to objects.

Although measurements from condition monitoring systems may be available there is still a challenge for the condition DT to respond on:

- Early warnings or anomaly detection
- Diagnostics, i.e., clarify which part is degraded, what is the degradation mechanism etc, see Figure 3.1
- Prognostics, i.e., how will degradation evolve over time, and when will the item not be able to perform it's required function any more, also see Figure 3.1 for *prediction*.

Typical approaches for handling these challenges are signal processing, first principle approaches (white box), probabilistic modelling and signal processing (grey box) and machine learning (black box).

3.5 Risk DT

For the operation and maintenance phase of a system the (static) risk models could be categorized into:

- Qualitative and semi-quantitative models and techniques like bow-tie, FMECA, HAZOP, Task Analysis and preliminary hazard analysis (PHA)
- Quantitative system risk and reliability models like fault tree analysis, event tree analysis and Markov analysis
- Structural reliability models including assessment of loads and strengths.

In order to make what-if inquiries to the risk DT it is important to explicit link the risk models to:

- The items and elements in the condition DT
- The environment DT
- Other factors that have an influence, and are expected to change over time, e.g., changes in maintenance resources.

3.6 Environment DT

The environment DT shall contain all aspects of the environment, this means for example in a real-time perspective current temperature, precipitation, wave heights and wind speed are important. But these dimensions can only partly cover the environmental impact on the asset. For example the impact of precipitation on a railway system in terms of risk of insufficient drainage capacity depends on existing water saturation of the soil, snow melting in surrounding areas etc. To establish relevant and useful environment DTs is therefore very demanding. An important element of an environment DT is weather forecasting.

3.7 Maintenance cost DT

The maintenance DT contains information regarding planned and executed maintenance, as well as logistics, personnel plans, how maintenance is organized etc. Cost models are required in order to set up what-if inquires for example related to changing time and amount of maintenance.

3.7.1 White-, grey- and blackbox models

The models introduced above are often denoted grey-box models because the degradation is only partial described, this in contrast to so-called white-box models. The following characteristics are often used

- *White-box models*. These are models where the physical degradation of an item is described by the laws of physics, chemistry etc. For example a fatigue model for crack propagation.
- *Gray-box models*. These are models where degradation is modelled by probabilistic models. In the current work Markov state models have been used, but other common models are the Gamma, Wiener and Inverse Gauss processes. Typically the degradation increments are described by stochastic jumps. These processes may include covariates representing physical conditions, but the relation between the increments and the physical conditions and factors are usually established by regression methods rather than physical laws. In Chapter 5 several probabilistic models are introduced.
- *Black-box models*. The black-box models aim to make predictions regarding future degradation without specifying any model. These models are "trained" by large amount of data. Typical models are machine learning methods like deep neural networks and Random forests.

In this course we primarily use gray-box models where some of these models are motivated by white-box models. For example there are physical degradation models stating that the degradation rate increases with increasing degradation level, which is then used when setting up the probabilistic gray-box model

Chapter 4

Preventive Maintenance

4.1 Introduction

The objective of this chapter is to demonstrate aspects of maintenance as part of a reliability analysis. In addition to passively treat maintenance as part of the reliability, we will also investigate some models for maintenance optimization.

☞ **Maintenance**: The combination of all technical and management actions during the life cycle of an item intended to retain the item in, or restore it to, a state in which it can perform as required.

Maintenance is important to achieve a high availability. Generally availability depends on the following factors:

1. Inherent reliability (e.g., quality, type of material used and design principles)
2. Maintainability (how easy it is to perform maintenance)
3. Maintenance support (resources, spare parts etc)

4.1.1 Maintenance Categories

The maintenance is often categorized into (Rausand et al., 2021):

1. Corrective maintenance (CM), i.e., tasks performed as a result of a detected item failure or fault, to restore the item to a specific condition. CM tasks may be carried out *immediately* or be *deferred*.

2. Preventive maintenance (PM), i.e., planned maintenance tasks performed prior to failures. The activities are carried out in order to reduce the probability of failure, or increase the mean time to failure (MTTF). There are several types of PM tasks:

   (a) Age-based

   (b) Clock-based (calendar based)

   (c) Condition-based

   (d) Opportunity-based

   (e) Overhaul, e.g., as part of a turnaround

3. Predictive maintenance, i.e., maintenance based on prognoses for the degradation of the item.

Note that the categorization varies from standard to standard, e.g., some standards include predictive maintenance as part of condition-based maintenance. Figure 4.1 depicts the categorization used in EN 13306.

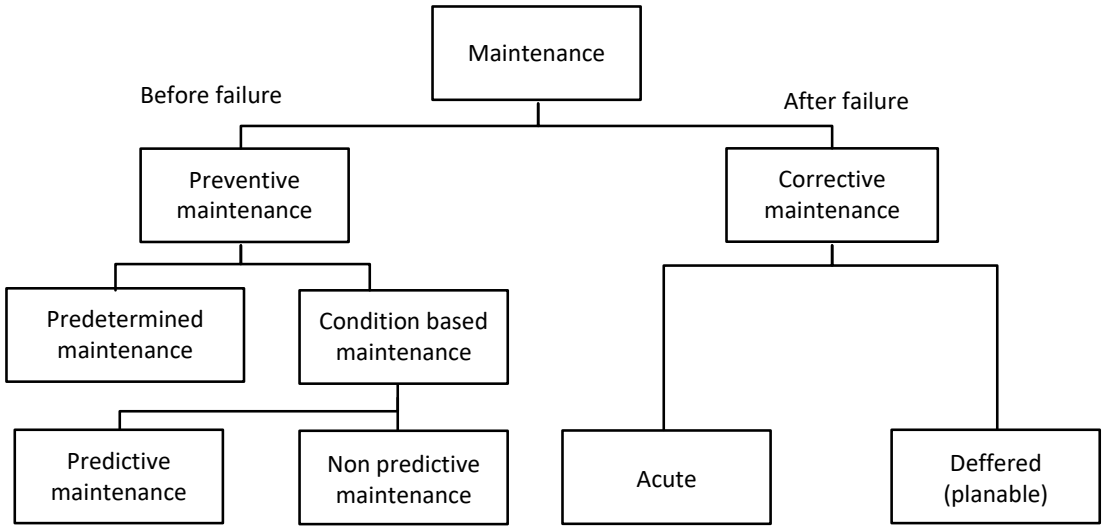


Figure 4.1: Maintennace categories (EN 13306)

4.1.2 Preventive Maintenance Policies

A preventive maintenance policy is a strategy that aims at minimizing the long run cost. A policy both deals with qualitative issues like replace an item periodically at a given age, and quantitative issues like what age that should be. The classical maintenance policies were basically considering age or calendar time as the decision variable to use in the optimization. In light of

"predictive maintenance" the condition of an item and future operational loads are becoming more important in order to minimize long run cost. Examples of both types of models will be investigated.

☞ **Preventive Maintenance**: Maintenance carried out at predetermined intervals or according to prescribed criteria and intended to reduce the probability of failure or functional degradation of an item.

4.1.3 Terminology and Cost Function

This section presents important terminology adapted from Rausand et al. (2021) with some additional terms.

- *Maintenance task:* A specific task to maintain an item determined by "what, where, how and when". A task is part of the task space, $\mathscr{A}$, i.e., $\mathscr{A} = a_1, a_2, a_3, \ldots$
- *Maintenance decision:* A process δ to select a specific maintenance task $a_i \in \mathscr{A}$. δ depends on available data $\mathscr{D}$, cost, operating conditions etc.
- *Maintenance strategy:* An overall framework describing how the maintenance decision problem shall be approached. A strategy embraces an objective function, often denoted the cost function:
- *Cost Function:* $C = C(a, \delta, t, \mathscr{D}, \mathscr{D}_{\text{OC}}, t_{\text{cal}}, \ldots)$. In addition to the maintenance task and the data the cost function depends on the time t of executing the maintenance, the operational context $\mathscr{D}_{\text{OC}}$, the calendar time $t_{\text{cal}}, \ldots$ (e.g., inside / outside working hours) and so on. A specific note is made regarding the notation used for the time. In some presentations t is used for the time axis, but in many other presentations we use τ to denote time, for example the length of a maintenance interval. When we deal with maintenance grouping we need to distinguish between the running time t and the local time x_i for the individual components. Also note that time may be multi-dimensional, for example if we carry out both a failure-finding-task and a replacement-task.

To optimize maintenance we would like to minimize the cost per unit time. In many situations this will be to minimize the expected maintenance cost in a renewal period divided by the expected length of the renewal period:

$$C = \frac{\mathrm{E}[C(T_R)]}{\mathrm{E}[T_R]} \tag{4.1}$$

The cost function in Equation (4.1) does not indicate any decision variable, i.e., which variables the cost function should be minimized with respect to. In literature the cost function is often denoted the *objective function*. The following situations are the most common:

- An item is periodically maintained at intervals of length τ. The cost function is denoted $C(\tau)$ and the challenge is to find the value of τ that minimizes the long run cost per unit time. The optimal value is denoted τ^*, and the minimal value of the cost function is denoted C^*.

- A situation has occurred where we need to make a "here-and-now" decision regarding the next maintenance. Current time is t_0 and running time from t_0 is denoted t. The time elapsed since last maintenance is x, i.e., maintenance was carried out at time $t_0 - x$. The objective is to determine time from t_0 until next maintenance is to be carried out. The cost function is denoted $C(t)$ and the objective is to find the value of t that minimizes the cost function over some limited time horizon. In this situation we usually use a so-called marginal cost approach.

- Also here we are going to make a "here-and-now" decision, but there are only a limited set of opportunities for maintenance. The first opportunity is now, i.e., at time t_0, the next opportunity is at time t_1 and there could be more opportunities at times $t_2, t_3, \ldots$ and so on. In this situation it is often not possible to specify an explicit objective function, hence we denoted the objective function $C_{t_0}, C_{t_1}, C_{t_2}, \ldots$ for the maintenance opportunities at times $t_0, t_1, t_2, \ldots$ respectively.

- The condition of an item is the critical information used to determine the next maintenance. If we are able to specify the condition by a one-dimensional health indicator, say $X(t)$ and assuming the item will fail when the health indicator exceeds a threshold ℓ, i.e., when $X(t) \geq \ell$, then it is reasonable to carry out maintenance whenever $X(t) \geq m$, where $m < \ell$. The cost function is then a function of the maintenance limit m, i.e., $C(m)$, and the objective is to find the value of m that minimizes $C(m)$. For items where the health indicator only can be revealed by inspections, we also need to determine the optimal inspection strategy, i.e., the cost function, $C(\tau, m)$ is depending both on the inspection interval τ and the maintenance limit m.

4.1.4 Reliability terminology

To develop maintenance models that can support preventive maintenance strategies requires a proper understanding of basic reliability terminology and models. In this section some basic ideas are presented, and readers familiar with reliability theory can skip this section. Reliability is about the ability to perform one or more required functions and we need some definitions:

☞ **Function**: An activity, process, or transformation stated by a verb and a noun that describes what must be accomplished.

Examples of functions are provide torch, stop flow of fluid and detect gas. Further:

☞ **Reliability**: The ability of an item to perform as required in a stated operating context and for a stated period of time.

Note 1: The term *item* could be a technical system a subsystem or a component

Note 2: The *required performance* must be specified, by e.g., laws, customer requirements etc.

Note 3: When we describe past reliability we use the term *achieved reliability*, whereas the single word *reliability* is always used to describe future reliability.
In some situation we will define reliability of a service, e.g., a bus service in a town:

☞ **Service reliability**: The ability of the service to meet its supply function with the required quality under stated conditions for a specified period of time.

4.1.5 States and transitions

Some items only operate in one state, e.g., a cooling pump may always be pumping. Other components operate in two or more states, e.g., a safety valve may be in an open position, or a closed position. For each state the item might have different functions. For example a valve in an open position has two main functions, i.e., keep open and close upon demand. Failing to perform a function is denoted a failure, and more precisely:

☞ **Failure**: The termination of the ability of an item to perform as required.

A failure is then an *event* that occurs in time, whereas a fault is a *state* where the item is not able to perform as required. An error is a “discrepancy between a computed, observed or measured value or condition and the true, specified or theoretically correct value or condition”. See Figure 4.2 for an illustration.

☞ **Fault of an item**: A state of an item, where the item is not able to perform as required.

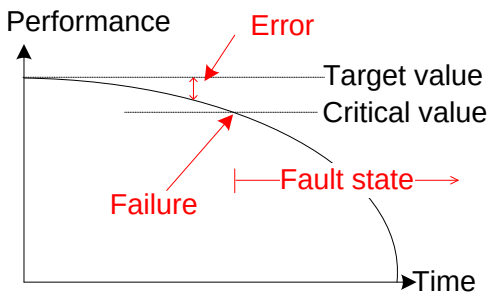


Figure 4.2: Failure and fault

☞ **Failure mode**: The manner in which a failure occurs, independent of the cause of the failure.

To understand the failure mode concept it is important to have focus on how the failure manifest it self, and not on the cause of the failure.

4.1.6 Failure causes and effects

☞ **Failure cause**: Set of circumstances that leads to failure.

The term 'cause' is a difficult term and in our context we distinguish between the "direct" or "proximate" cause and the "root cause", i.e.,:

☞ **Proximate cause**: An event that occurred, or a condition that existed immediately before the failure occurred, and, if eliminated or modified, would have prevented the failure.

☞ **Root cause**: One of multiple factors (events, conditions, or organizational factors) that contributed to or created the proximate cause and subsequent failure and, if eliminated, or modified would have prevented the failure.

Note that the direct or proximate cause on one level in a system hierarchy may be the effect of a failure mode on a lower level. For example the proximate failure cause of a pump might be a "bearing failure", where again the failure mode of the bearing is to provide "friction less" rotation of the impeller.

"Behind" the bearing failure we may find the root cause, e.g., lack of lubrication (grease). To trace root causes we may even go further behind, e.g., lack of maintenance and even deficiency in the maintenance management.

4.1.7 State variable

The state variable of an item is used to specify the state of a component. In some situations we use the state variable to enumerate the various state a component can be in, i.e., x, whereas in other situations we will also treat the stochastic behaviour of the item, hence the state variable will be a stochastic variable:

$$X(t) = \begin{cases} 1 \text{ if the item is functioning at time } t \\ 0 \text{ if the item is in a failed state at time } t \end{cases} \tag{4.2}$$

Note that we use an uppercase letter for the state variable when treated as a stochastic variable, and a lowercase letter when we just enumerate or consider a specific value of the state of the item.

4.1.8 Time-to-failure

The *time-to-failure*, or *lifetime* of an item is the time elapsing from when the item is put into operation until it fails for the first time. If we denote the time-to-failure with T then

$$T = \min\{t : X(t) = 0\}$$

Note that "time" sometimes is measured indirectly, e.g., by the number of kilometres driven by a car, the number of times a switch is operated, and the number of rotations of a bearing.

Since $X(t)$ is a stochastic variable, the time-to-failure, T, is also a stochastic variable. To grasp the reliability metrics we could relate these metrics to what we would observe if did experiments and collected the true lifetimes of the items. In the textbook several examples of such "empirical metrics" are given.

4.1.9 PDF and CDF

Assume that the time-to-failure T is a continuous distributed stochastic variable with probability density function $f(t)$ and cumulative distribution function $F(t)$. Figure 4.3 shows the probability density function (PDF).

To interpret the PDF we have for small Δt:

$$\Pr(t < T \leq t + \Delta t) \approx f(t)\Delta t$$

i.e., the probability that a new item will fail in the interval t to $t + \Delta t$ equals the PDF at time t multiplied with the length of the interval.

Figure 4.3 does not give any indication why the item fails. A failure mechanism is a physical

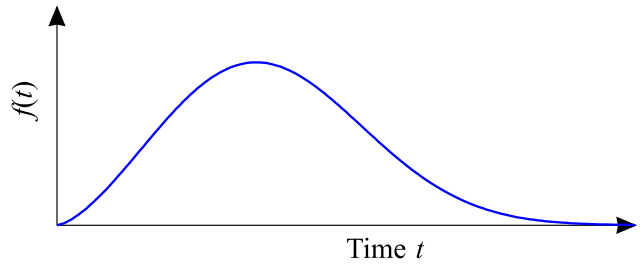


Figure 4.3: Probability density function, $f(t)$

or chemical process that leads to failure. Fatigue is one such mechanism where fatigue cracks develop into a breakage, i.e., a failure. Figure 4.4 illustrates the situation.

The lower part of the figure illustrates the crack propagation. Due to different loads the crack propagation is considered as a stochastic process, and different trajectories are indicated. When the crack size reach a critical value, i.e., the failure limit l in the figure, the item will fail. Since the crack propagation is random, also the time-to-failure will be random, and the corresponding probability density function is indicated in the upper part of the figure.

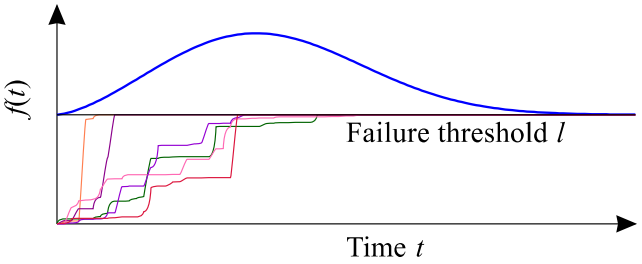


Figure 4.4: Different trajectories of crack propagation leads to stochastic time-to-failure

In some cases it is possible to measure or by other means assess the underlying process. In such cases we know a specific trajectory as time goes by, and we may utilize this knowledge to make more precise prediction of the time-to-failure as time goes by, that is $f(t)$ in Figure 4.4 will be sharper and sharper. If the underlying development cannot be observed, the original $f(t)$ in Figure 4.3 is the only knowledge we have regarding coming failures. Section 4.1.11 introduces the failure rate function, which is the *conditional* probability of failure as a function of time, and will be the expression to use in order to determine when to perform a preventive maintenance action.

The relation between the cumulative distribution function (CDF) and the probability density

function is given by:

$$F(t) = \Pr(T \leq t) = \int_0^t f(u)du$$

4.1.10 Survivor Function

The survivor function of an item is defined by:

$$R(t) = 1 - F(t) = \Pr(T > t) = \int_t^\infty f(u)du$$

i.e., the probability that a new item will survive the time interval $(0, t]$.

4.1.11 Failure Rate Function

The failure rate function is essentially the conditional probability that an item will fail in a small time interval given that it has not failed up till now. The probability that an item will fail in $(t, t+\Delta t]$ when we know that the item is functioning at time t is:

$$p(t,\Delta t) = \Pr(t < T \leq t+\Delta t | T > t) = \frac{\Pr(t < T \leq t+\Delta t)}{\Pr(T > t)} = \frac{F(t+\Delta t) - F(t)}{R(t)}$$

If we investigate the ratio $p(t,\Delta t)/\Delta t$ we get the failure rate function $z(t)$ of the item:

$$\begin{aligned} z(t) = \lim_{\Delta t \to 0} \frac{p(t,\Delta t)}{\Delta t} &= \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t+\Delta t | T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} \frac{1}{R(t)} = \frac{f(t)}{R(t)} \end{aligned}$$

And for small Δt:

$$\Pr(t < T \leq t+\Delta t | T > t) \approx z(t)\Delta t$$

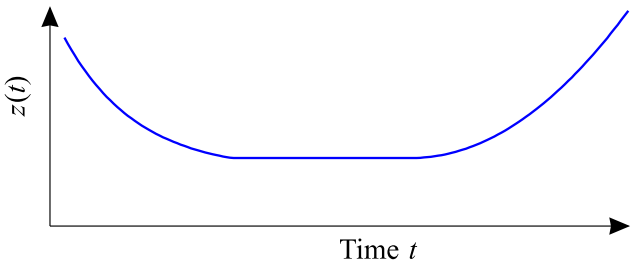


Figure 4.5: Typical shape of $z(t)$, i.e., a bathtub shape

Figure 4.5 shows a typical shape of $z(t)$. This shape of the failure rate function is the origin of the name *bathtub curve* for the failure rate function. In an early life the item may be exposed to so-called "burn-in" failures. This could be failures related to errors during installation, or defects during manufacturing. Then there could be a rather long period of normal operation with low probability of failure. As time goes by "wear-out" often brings the item to it's end of life. It should be noted that not all items will follow the bathtub curve.

Note that $f(t)$ is the probability of failing at time t, whereas $R(t)$ is the probability of surviving time t. In $z(t) = \frac{f(t)}{R(t)}$ we divided by $R(t)$, so even if the probability of failing at large times t is low, the fraction becomes very high since we hardly survive t, i.e., the denominator $R(t) \approx 0$.

4.1.12 Effective failure rate

The effective failure rate, $\lambda_E(\tau)$, is the unconditional expected number of failures per time unit as a function of the maintenance interval τ. Consider an item with an increasing failure rate function $z(t)$ at the end of life:

- If no preventive maintenance is conducted and the item is only maintained upon a failure, the effective failure rate will be quite high, see the left part of Figure 4.6
- If we maintain at time τ_1 this corresponds to "removing" a part of the right hand side of the bathtub curve, resulting in a lower effective failure rate, i.e., shown in the middle of Figure 4.6
- If we further reduce the maintenance interval, say to τ_2 the resulting effective failure rate could be quite low, i.e., shown in the right part of Figure 4.6

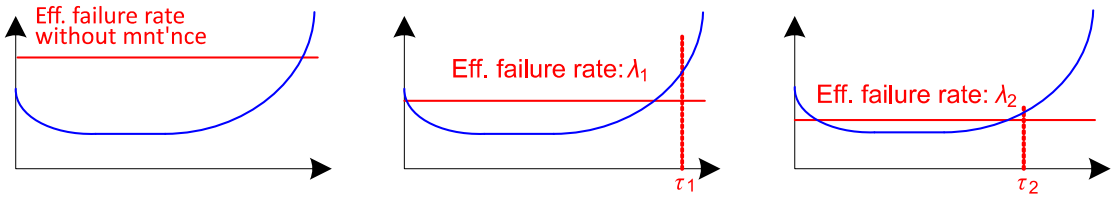


Figure 4.6: Effective failure rate for various maintenance decisions

If the failure rate function is known we may approximate the effective failure rate by the average failure rate function in the interval $[0, \tau)$:

$$\lambda_E(\tau) \approx \frac{1}{\tau} \int_0^\tau z(t)\,dt \tag{4.3}$$

4.1.13 Age and calendar based policies

The age and block replacement policies are two classical maintenance models that can motivate the models we will derive in this course.

Age Replacement Policy - ARP

In the age replacement policy an item is replaced or overhauled to an as-good-as-new condition when the item reaches a specified age. We usually consider a replacement rather than an overhaul, but the situation is the same if an overhauled item is as-good-as-new after an overhaul. The age at replacement is denoted τ and the challenge is to find the optimal τ, say τ^*. The situation is characterized by:

- The item is replaced with a new item, or repaired to an as-good-as-new (AGAN) state when it reaches the age τ
- If the item fails before the scheduled maintenance, the unit is replaced and the "maintenance clock" is set to 0
- In Figure 4.7 T_1 and T_2 are failure times where the item is replaced with a new item or repaired to an as-good-as-new state
- The cost of a preventive replacement is c
- The cost of a corrective replacement, i.e., replacing a failed item is $c + k$

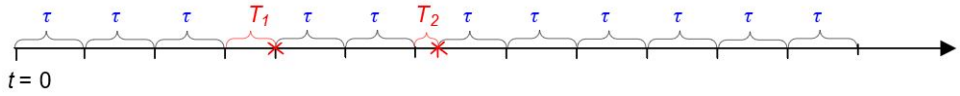


Figure 4.7: ARP

Let $f(t)$ denote the time-to-failure distribution of the item and assume that the item is as-good-as-new after a replacement. The time between two consecutive replacements is called a replacement period. This period is stochastic, and the mean time between replacements is:

$$\text{MTBR}(\tau) = \int_0^\tau t f(t) dt + \tau \Pr(T > \tau) = \ldots = \int_0^\tau (1 - F(t))\, dt = \int_0^\tau R(t) dt$$

where T is the (potential) time to failure, and we have used partial integration to derive the formula.

For each replacement period we always have to pay the cost c. If a replacement period ends with a failure, we have to pay an extra cost k. The probability of paying the extra cost is $\Pr(T \leq$

$\tau) = F(\tau)$. The long run cost per unit time is then given by:

$$C_{\mathrm{A}}(\tau) = \frac{\text{Cost in a cycle}}{\text{Expected length of a cycle}} = \frac{c + kF(\tau)}{\int_0^\tau (1 - F(t))\, dt}$$

Numerical methods are required to minimize $C_{\mathrm{A}}(\tau)$

Numerical methods for calculating MTBR(τ)

In case of Weibull distributed times-to-failure, the following Python code may be used to calculate MTBR(τ):

```
from scipy.integrate import quad
import math
def iMTBR(t,alpha,lmbda):
    # Integrand, i.e., 1-F(t;alpha,lmbda)
    return math.exp(-((t*lmbda)**alpha))

def MTBR(tau, MTTF, alpha):
    lmbda=math.gamma(1+1/alpha)/MTTF
    I,err = quad(iMTBR, 0, tau, args=(alpha, lmbda))
    return I

print("Test: MTBR",MTBR(40,100,3))
```

where `iMTBR()` is a function returning the integrand in $\int_0^\tau (1 - F(t))\, dt$. To minimize the objective function in Equation (4.4) we can make a plot in Python:

```
from scipy.integrate import quad
from numpy import arange, zeros
import math
import matplotlib.pyplot as plt
def iMTBR(t,alpha,lmbda):
    # Integrand, i.e., 1-F(t;alpha,lmbda)
    return math.exp(-((t*lmbda)**alpha))
def MTBR(tau, MTTF, alpha):
    # Carry out numerical integration by using scipy quad-function
    lmbda=math.gamma(1+1/alpha)/MTTF
    I,err = quad(iMTBR, 0, tau, args=(alpha, lmbda))
    return I
def C_A(tau, MTTF, alpha, c, k):
    # Objective functino
    lmbda=math.gamma(1+1/alpha)/MTTF
    return (c + k*(1-math.exp(-((tau*lmbda)**alpha))))/ \
        MTBR(tau, MTTF, alpha)
# Parameters
```

```
MTTF = 4
alpha = 3
c = 15
k = 51
xlist=zeros([21])
ylist=zeros([21])
# Calculate objective function for relevant arguments
i=0
for tau in arange(1,3.1,0.1):
    xlist[i]=tau
    ylist[i]=C_A(tau,MTTF,alpha,c,k)
    i+=1
plt.plot(xlist, ylist)
plt.xlabel(r'$\tau$')
plt.ylabel(r'$C_A(\tau)$')
plt.title('Cost as function of maintenance interval')
plt.savefig('ARPoutput.svg')
plt.show()
```

Block Replacement Policy - BRP

In a block replacement policy an item is periodically replaced at predefined points of time. The argument for this could be that we have many identical components, and it is more convenient to perform the preventive maintenance at the same time (i.e., a block replacement). Another argument for a block replacement policy could be that this is much easier to manage by our computerized maintenance management system (CMMS). Figure 4.8 illustrates the situation. T_1 and T_2 are failure times, but they will not affect the time of the next preventive activity. It seems a bit "waste" of useful life to replace at time 4τ but this may be defended by lower administrative cost.

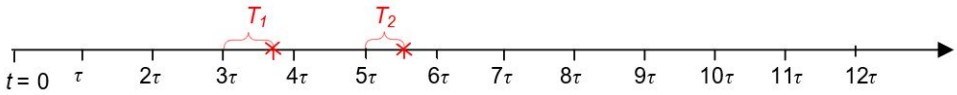


Figure 4.8: BRP

The situation is slightly different from the ARP, and also a different cost structure is used:

- The item is replaced every τ time unit
- An item may fail one ore more times between periodic replacements, if this happens it it assumed that the item is immediately replaced or repaired to a-good-as-new condition
- The cost of a preventive replacement is c

- The cost of a corrective replacement, i.e., replacing a failed item is k

In this situation the replacement period is always τ. In each period we always have to pay the cost c. In addition we pay the cost k for each failure. The expected number of failures is given by the *renewal function*, $W(\tau) = \mathrm{E}[N(\tau)]$. Thus the average cost per time unit is:

$$C_B(\tau) = \frac{c + kW(\tau)}{\tau}$$

Note that $W(\tau)/\tau$ is the average expected number of failures per time unit when the item is replaced every τ time unit. This is often written:

$$\lambda_{\mathrm{E}}(\tau) = \frac{W(\tau)}{\tau}$$

For small values of τ compared to the MTTF, it is unlikely that we have more than one failure in an replacement period. This means that the expected number of failures in a replacement period is given by the average value of the failure rate function, $z(t)$. If we assume that failure times are Weibull distributed, we obtain:

$$\lambda_{\mathrm{E}}(\tau) = \left(\frac{\Gamma(1+1/\alpha)}{\mathrm{MTTF}}\right)^{\alpha} \tau^{\alpha-1}$$

where $\mathrm{MTTF} = \Gamma(1+1/\alpha)/\lambda$. In this situation it is straight forward to find an analytical solution for the optimal replacement period:

$$\tau^* = \frac{\mathrm{MTTF}}{\Gamma(1+1/\alpha)} \sqrt[\alpha]{\frac{c}{(\alpha-1)k}}$$

4.1.14 Marginal cost approach

The age- and block replacement policies discussed in the previous section find the optimal interval by minimizing expected cost per time unit. This is the standard approach we use to optimize maintenance intervals. However, this approach is rather static and can not take into account real-time information that would be relevant for optimizing the next maintenance interval.

To include here-and-now information relevant for the next maintenance we often use a marginal cost approach. That is, we consider the situation from now on until the next preventive or corrective maintenance action. Let t be running time from now on, and let x be the time elapsed since last preventive maintenance action, i.e., the "local age" of the item under consideration. We now seek the value of t that minimizes the (marginal) expected cost in the interval $[0, t]$ plus the average cost from t until the end of a longer time horizon to consider, i.e., up to some time T.

In order to obtain the average cost in the longer perspective, assume that we already applied the ARP or BRP approach to calculate some average cost per unit time, say C^*. The challenge is now to calculate the expected cost in the interval $[0, t]$. Various aspects could be taken into account. To demonstrate the approach, we consider the age replacement policy, and assume that the cost of the preventive activity c depends on the time t, i.e., $c = c(t)$, but the additional failure cost k is fixed. The marginal cost approach is now to minimize the objective function:

$$C(t) = kF(t|x) + C^*(T-t) + C^* \int_0^t (t-u) f(t|x) du + c(t) \tag{4.4}$$

where $F(t|x) = 1 - R(t+x)/R(t)$ and $f(t|x) = f(t+x)/R(t)$ are the conditional CDF and PDF respectively for time-to-failure given that the item has survived up to time x. Further $R(t)$ and $f(t)$ are the unconditional survivor function and probability density function respectively.

Note that to minimize the objective function in (4.4) we can choose any T which is larger than the optimal t. So if we have some tentative optimal value, we could just let T be 10 times longer. Further we need the average cost per time unit, C^*.

4.1.15 Interval optimization - General approach for non-observable failure progression

Reliability Centred Maintenance (RCM) is a systematic approach to determine appropriate maintenance tasks. Chapter 13 presents the main steps required to run an RCM exercise. RCM provides a dedicated decision logic to determine the type of maintenance, and we essentially distinguish between the following situations:

1. It is possible to observe a (health) indicator that can warn about coming failures, see Figure 4.4
2. It is not possible observe such an indicator, but there are ageing mechanisms, see Figure 4.3
3. There is no indicator, there is no ageing, and the function of the item is hidden.

In this section we focus on situation 2 where we do not have any indicator, but we may use the age of the component, the number of hours operated, the total mileage run by a car etc. to decide upon next preventive maintenance.

Consider an item where a preventive maintenance (PM) action is conducted at predetermined intervals due to an increasing failure rate function $z(t)$. Typically we assume that time-to-failure is Weibull distributed where the failure rate function is given by:

$$z(t) = \alpha \lambda^\alpha t^{\alpha-1} \tag{4.5}$$

In order to find an optimal interval for the preventive maintenance action we establish the average cost per time unit as a function of the maintenance interval, say τ:

$$C(\tau) = c_{\text{PM}}/\tau + \lambda_{\text{E}}(\tau)\,[c_{\text{CM}} + c_{\text{EP}} + c_{\text{ES}}] \tag{4.6}$$

where c_{PM} is the cost of a preventive maintenance action (to prevent failures), c_{CM} is the cost of a corrective maintenance (CM) action if a failure occurs, $\lambda_{\text{E}}(\tau)$ is the *effective failure rate*, i.e., the expected number of failures per time unit when the component is preventively maintained every τ time unit, c_{EP} is the expected production losses upon a component failure, and finally c_{ES} is the expected safety cost upon a component failure, including material damages and environmental losses.

There is no general formula for obtaining c_{EP} and c_{ES} upon an item failure. Within the domain of safety and reliability analysis there are several models and methods that apply. More information about these methods can be found in the course ntnu.no/studier/emner/PK6031.

Often we are able to make a direct argument to establish c_{EP} and c_{ES}, and for the production losses we are often able to specify the cost by:

$$c_{\text{EP}} = p_{\text{P}}\,(c_{\text{D}}\text{MDT} + c_{\text{T}}) \tag{4.7}$$

where p_{P} is the probability that a failure of the actual item results in a production loss, c_{D} is the production loss per hour (in NOKs or Euros) given a system failure, MDT is the mean down time upon a failure and c_{T} is the trip cost, i.e., a cost that is paid, but independent of the duration of the downtime.

Often the mean down time is split into MDT = MLD + MRT where MLD is the mean logistic delay time, and MRT is the mean active repair time. For offshore wind MLD would usually be the dominating factor.

To minimize the objective function in Equation (4.6), we usually let $c_U = c_{\text{CM}} + c_{\text{EP}} + c_{\text{ES}}$ denote the expected unplanned cost upon a failure to simplify.

The effective failure rate depends on the time-to-failure distribution of the item. The Weibull distribution is a widely used distribution for ageing components. In the case of Weibull distributed times-to-failure we may find approximation formulas for the effective failure rate. If we know the mean time to failure, MTTF (without maintenance), and the ageing parameter, α, of the time-to-failure distribution of the item, the effective failure rate may be approximated by: eqStream: Effective failure rate approximation:

$$\lambda_{\text{E}}(\tau) = \left(\frac{\Gamma(1+1/\alpha)}{\text{MTTF}}\right)^{\alpha} \tau^{\alpha-1} \tag{4.8}$$

where $\Gamma(\cdot)$ is the gamma function. The approximation is good when the maintenance interval

is small compared to the MTTF. If the maintenance interval is approaching the MTTF value, the approximation in Equation (4.8) is not very accurate for large values of τ, and we might use the following improved approximation, see Kwang Pil et al. (2008):

$$\lambda_{\mathrm{E}}(\tau) = \left(\frac{\Gamma(1+1/\alpha)}{\mathrm{MTTF}}\right)^{\alpha} \tau^{\alpha-1}\left[1 - \frac{0.1\alpha\tau^2}{\mathrm{MTTF}^2} + \frac{(0.09\alpha - 0.2)\tau}{\mathrm{MTTF}}\right] \tag{4.9}$$

The approximated effective failure rate in Equation (4.9) is usually sufficient, but if higher precision is required we may use renewal theory. From the fundamental renewal equation we have $W(t) = F_T(t) + \int_0^t W(t-x) f_T(x) dx$. Here $W(t)$ is the expected number of events up to time t in a renewal process. In case we have a reasonable initial numerical approximation for $W(t)$, say $W_0(t)$ we may use the following iteration scheme:

$$W_i(t) = F_T(t) + \int_0^t W_{i-1}(t-x) f_T(x) dx$$

to obtain better and better solutions for $W(t)$. As a starting point we use:

$$W_0(t) = \lambda_{\mathrm{E}}(t)\, t = \left(\frac{\Gamma(1+1/\alpha)}{\mathrm{MTTF}}\right)^{\alpha} t^{\alpha-1} t = \left(\frac{\Gamma(1+1/\alpha)}{\mathrm{MTTF}}\right)^{\alpha} t^{\alpha}$$

In the following we assume that the approximation in Equation (4.8) is sufficient for our purpose. By equating the derivative of $C(\tau)$ in Equation (4.6) to zero, we find the optimal interval to be:

$$\tau^* = \frac{\mathrm{MTTF}}{\Gamma(1+1/\alpha)} \left(\frac{c_{\mathrm{PM}}}{c_{\mathrm{U}}(\alpha-1)}\right)^{1/\alpha} \tag{4.10}$$

Example 4.1 Wind turbine

We are considering a wind turbine of 10 MW. In average we assume that the output effect is 6 MW, where wake effects, periods of low wind speed etc. cause reduction in the produced energy. The average loss per kilowatt hours (kwh) is assumed to be 0.5 NOKs. An electrical motor is used for jawing control. In case of a failure of the motor we assume that we have to shut down the wind turbine until the motor is repaired or replaced.

The motor is assumed to have an increasing failure rate function, where the ageing is rather strong, i.e., the ageing parameter is $\alpha = 3$.

MTTF is assumed to be five years if no preventive maintenance is carried out. Under normal weather conditions we the mean downtime is 12 hours. The cost of a preventive replacement of the motor is $c_{\mathrm{PM}} = 15\,000$ NOKs. The cost of a corrective replacement is $c_{\mathrm{CM}} = 30\,000$ NOKs. In addition to the cost of repairing the failed motor, there is a loss of electricity production.

In order to apply Equation (4.10) we need the PM cost, i.e., $c_{\mathrm{PM}} = 15\,000$ NOKs, the CM cost,

i.e., $c_{CM} = 30\,000$ NOKs, and the downtime cost, i.e., $c_{EP} = \text{MDT} \cdot 0.5 \cdot 6\,000 = 36\,000$ NOKs. Total (unplanned) cost upon a failure is thus $c_U = c_{CM} + c_{EP} = 66\,000$ NOKs. Note that in this example $p_P = 1$, and $c_T = 0$. The optimal interval is given by:

$$\tau^* = \frac{\text{MTTF}}{\Gamma(1+1/\alpha)} \left(\frac{c_{PM}}{c_U(\alpha-1)} \right)^{1/\alpha} = \frac{4}{\Gamma(1+1/3)} \left(\frac{15\,000}{66\,000 \cdot 3} \right)^{1/3} \approx 2.7 \text{ years}$$

Alternatively, we could solve the problem by a minimization routine, for example by the Solver in Excel. The cost function to minimize is:

$$C(\tau) = c_{PM}/\tau + \lambda_E(\tau) c_U$$

□

Often we introduced the pre-calculated effective failure rates:

- Low ageing: $\lambda_E(\tau) = 0.79\tau/\text{MTTF}^2$
- Medium ageing: $\lambda_E(\tau) = 0.71\tau^2/\text{MTTF}^3$
- Strong ageing: $\lambda_E(\tau) = 0.67\tau^3/\text{MTTF}^4$

Since $\alpha = 3$ we may alternatively use the formula for medium ageing when solving the problem in e.g., Excel.

As a third option, we could also make a plot of the cost function. In this case we might visualize all cost elements, i.e., we use

$$C(\tau) = c_{PM}/\tau + \lambda_E(\tau) c_U = c_{PM}/\tau + \lambda_E(\tau) (c_{CM} + c_{EP})$$

Figure 4.9 depicts the cost elements. Hours is the time unit on the x-axis, and the optimal value is around 24 000 which is slightly less than three years as found by the analytical solution. Note that expected production losses dominate the corrective maintenance cost.

4.1.16 Digital twin

The objective function in Equation (4.6) is used to find an optimal maintenance interval for items where we cannot utilize the condition of the item, but where there is ageing. The result is an interval which in the long run is optimal. In some cases, there are specific conditions which will affect our decision regarding maintenance. For example:

- The maintenance window is “closed” at the optimal point of time for maintenance

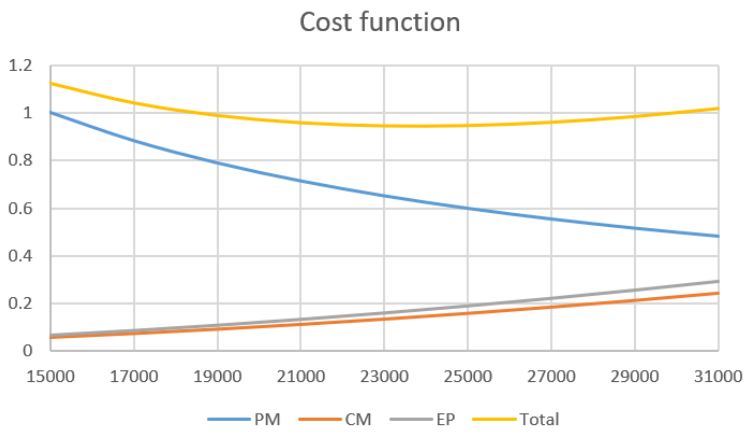


Figure 4.9: Cost function for the yaw motor

- The preventive maintenance cost is temporarily deviating from the average preventive maintenance
- Production losses upon a failure is temporarily deviating from the average value
- etc.

In principle we would have a "digital twin" for production, for the weather etc. which could be combined with the objective function in Equation (4.6) in order to make real-time decision regarding the point of time for the next maintenance. For example, if the price of electricity is very high, i.e., c_{EP}, and we are approaching the time of next maintenance, we could insert the current value of c_{EP} and minimize Equation (4.6) to find the next point of time for maintenance.

A more demanding situation is if maintenance window is "closed" for a period of time when preventive maintenance is scheduled. We will elaborate on this, and make the following assumption:

- Current time is t_0
- The current age of the item is x, i.e., it is x time unit since last maintenance was carried out, thus the last maintenance was then carried out at time $t_0 - x$
- τ^* is the optimal maintenance interval
- The maintenance window will close just after time is t_0, and reopen at time t_1, where $t_0 < t_0 - x + \tau^* < t_1$, i.e., the due date for next maintenance is within the closed maintenance window
- The alternatives for executing the preventive maintenance is therefore t_0 and t_1

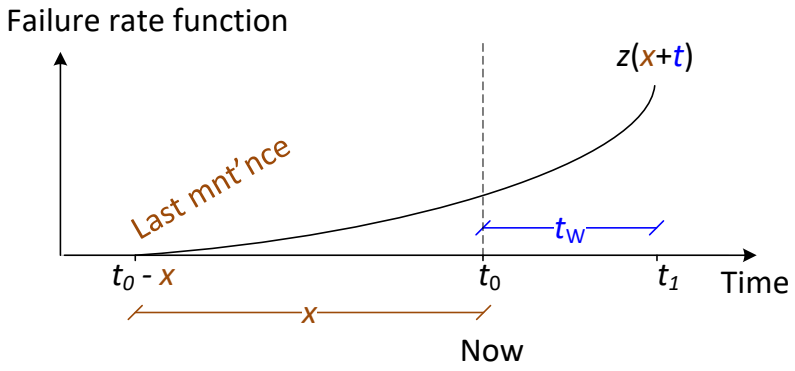


Figure 4.10: Failure rate function in relation to the weather window

- If we postpone the maintenance to t_1, and the item fails before the maintenance window opens, we will loose the production until the maintenance window opens.

Figure 4.10 depicts the failure rate function from $t_0 - x$ to t_1.

In the modelling we still assume that time-to-failure is Weibull distributed with PDF and survivor function given by:

$$f(t) = \alpha\lambda(\lambda t)^{\alpha-1} e^{-(\lambda t)^{\alpha}} \tag{4.11}$$

and

$$R(t) = \Pr(T > t) = 1 - F_T(t) = e^{-(\lambda t)^{\alpha}} \tag{4.12}$$

where

$$\lambda = \frac{\Gamma(1/\alpha + 1)}{\text{MTTF}} \tag{4.13}$$

In the modelling we assume there is a fixed cost of failure c_{T} and corrective cost C_{CM}. But in addition to the immediate cost of a failure during the period of a closed maintenance windo is c_D per time unit until the component could be repaired.

If we decide to carry out the PM activity at time t_0, i.e., just before the maintenance window closes the expected cost until the maintenance window opens at time t_1 is:

$$C'_{t_0} = c_{\text{PM}} + [1 - R(t_{\text{W}})]\,(c_{\text{CM}} + c_{\text{T}}) + c_D \int_{t=0}^{t_{\text{W}}} f(t)\,(t_{\text{W}} - t)\,dt \tag{4.14}$$

If we postpone the next PM activity to time t_1 the expected cost up to, but not including time t_1 is:

$$C'_{t_1} = [1 - R(t_W + x)/R(x)]\,(c_{CM} + c_T) + c_D \int_{t=0}^{t_W} f(t+x)\,(t_W - t)\,dt/R(x) \tag{4.15}$$

Note that C'_{t_1} is not including the preventive maintenance cost time t_1. Preventive maintenance cost is not to be paid if there has been a failure during the closed maintenance window, hence we add the expected preventive maintenance cost in order to obtain:

$$C_{t_1} = C'_{t_1} + c_{PM} R(t_W + x)/R(x) \tag{4.16}$$

Further if PM is carried out at time t_0 the time to the next preventive maintenance is t_W shorter compared to waiting until time t_1, hence to be able to compare we add the expected total cost in a time interval of length t_W, i.e.,

$$C_{t_0} = C'_{t_0} + C^* t_W \tag{4.17}$$

where $C^* = C(\tau^*)$ is given by Equation (4.6). The PM should be carried out at time t_0 if $C_{t_0} < C_{t_1}$, else we should wait until the maintenance window opens at time t_1.

4.2 Hidden function

If the function of a system being analysed is hidden, periodic functional or proof testing may reveal failures. In this situation, therefore, the maintenance activity is periodic proof test. The more frequent a proof test is carried out, the less likely the item will be in a fault state upon a demand. Such a demand could be to activate the process shut-down in emergency situations, or to start a back-up pump in case of the main pump is failing. In this situation we calculate the so-called PFD (Probability of Failure on Demand) indicating the proportion of time the system cannot perform the required function. Functions having hiden functions often comprises N identical elements and we require k or more of these elements to function in order to ensure that the system is functioning. Such systems are denoted a $k$oo$N$ system. The PFD is found by:

$$\text{PFD}(\tau) \approx \binom{N}{N-k+1} \frac{(\lambda\tau)^{N-k+1}}{N-k+2} + \beta\lambda\tau/2 \tag{4.18}$$

where

- β is the proportion of common cause failures, i.e., causes that result in a simoulatneous failure of all N elements
- If there is only one element, we use the usual formula PFD = $\lambda\tau/2$

- The binomial coefficient is given by $\binom{x}{y} = \frac{x!}{y!(x-y)!}$

The cost equation is given by:

$$C(\tau) = c_{\text{FT}}/\tau + N \cdot c_{\text{R}}\lambda(1 - \lambda\tau/2) + c_{\text{H}} \cdot \text{PFD}(\tau) \cdot f_D \tag{4.19}$$

where

- c_{FT} is the cost of performing a proof test of all the N elements
- c_{R} is the cost of repairing one of the N elements if found in a fault state by a proof test
- c_{H} is the cost of a hazardous event
- f_D is the rate of demands, e.g., the rate of gas leaks
- $\text{PFD}(\tau)$ is given by Equation (4.18)

In this situation, we cannot find a solution by equating the derivative of the cost equation (4.18) to 0, and we must then either minimize $C(\tau)$ numerically, or graphically.

To better understand the term $N \cdot c_{\text{R}}\lambda(1 - \lambda\tau/2)$ we realize that the mean time to failure is $1/\lambda$ but a failure is not revealed immediately, and in average it takes $\tau/2$ time units until a failure is revealed. This means that the mean cycle length, or mean time between failures, is $1/\lambda + \tau/2 = (1 + \lambda\tau/2)/\lambda$ and the corresponding *frequency* of repairs is $\lambda/(1 + \lambda\tau/2)$. Using the Taylor expansion $1/(1 + x) \approx 1 - x$ gives $\lambda/(1 + \lambda\tau/2) \approx \lambda(1 - \lambda\tau/2)$. Multiplying with the number of elements N and the repair cost C_{R} gives the final cost for repairs.

Problems

4.1 Timing Belt. The timing belt of a car is a critical component. If it fails, there is a large risk that this causes serious damages to the engine. In this problem we assume the following:

- $\text{MTTF}_{\text{WO}} = 175000$ km (WO means Without Maintenance, i.e., the MTTF if we do not replace the timing belt preventively)
- $\alpha = 3$ = ageing (shape) parameter in the Weibull distribution for time-to-failure
- $c_{\text{PM}} = 7000$ NOKs (Cost of preventively replacing the timing belt)
- $c_{\text{CM}} = 35000$ NOKs (Cost if the timing belt fails, i.e., major damages to the engine)

Find the optimal interval for replacing the timing belt by using the following methods:

1. Analytical, i.e., taking derivatives and set equal to 0

2. Graphical solution

3. Numerical solution (for example the Excel Solver or scipy.optimize in Python)

4.2 Timing Belt, continued. The timing belt Problem 4.1 is more realistic if we include the following assumptions:

- Pr(Need to rent a car|Breakdown) = 0.1
- Cost of renting a car = NOK 5000
- Pr(Overtaking |Breakdown) = 0.005
- Pr(Collision|Overtaking |Breakdown)=0.2
- $c_{\text{Collision}}$ = 25 million NOKs

Find the optimal interval in this situation, and compare with the previous exercise.

4.3 Jaw motor example Implement the jaw motor example in Python where you use a numerical routine for minimizing the objective function. Compare the result by using Equation (4.8) and Equation (4.9) for the effective failure rate.

4.4 Consider the motor used for jawing control above where the ageing parameter is $\alpha = 4$ and MTTF = 5 years if no preventive maintenance is carried out. The cost of a preventive replacement of the motor is c_{PM} = 15 000 NOKs. The cost of a corrective replacement is c_{CM} = 30 000 NOKs. The cost of loss production is $0.5 \cdot 6000 = 3000$ NOKs per hour. Assume that current age of the yaw motor is $t_0 = x = 17\,000$ hours, i.e., the due time for maintenance is approaching. But, it is expected difficult to approach the turbine for the next 14 days. That is the next opportunity for maintenance is $t_1 = t_0 + 14 \cdot 24 = 17\,336$. Determine if it pays off to advance the next PM. Note that t is used to denote running time, but here we can assume that $t = 0$ corresponds to the last maintenance, i.e., $x = 17\,000$ hours ago.

4.5 We are considering the maintenance of an emergency shutdown valve (ESDV). The ESDV has a hidden function, and it is considered appropriate to perform a proof test of the valve at regular intervals of length τ. The cost of performing such a test is NOK 10 000. If a test reveals a hidden failure the cost of repairing the failed valve equals NOK 50 000. If the ESDV is demanded in a critical situation, the total (accident) cost is NOK 10 000 000. The rate of demands for the ESDV is one every 5 year. The constant failure rate of the ESDV is $2 \cdot 10^{-6}$ hrs^{-1}). Determine the optimum value of τ by:

- Finding an analytical solution
- Plotting the total cost as a function of τ

- Minimising the cost function by means of numerical methods

4.6 In order to reduce testing it is proposed to install a redundant ESDV. The extra yearly cost of such an ESDV is NOK 15 000. Determine the optimum test interval if we assume that the second ESDV has the same failure rate as the first one, and there is a common cause failure situation, with $\beta = 0.1$. Will you recommend the installation of this redundant ESDV?

4.7 We will assess the maintenance of a pump system. The pump system consists of an active pump, pump A, and a stand-by pump, pump B. Typically, pump A is run during normal operation. If pump A fails, pump B can be started. If we succeed in starting pump B, we assume that pump B will not fail while pump A is being repaired. After pump A is repaired, pump A is put into operation, while pump B is put back into cold stand-by. For pump A, we assume that an overhaul will ensure that the pump is almost as good as new after the overhaul. Furthermore, a BRP strategy is followed. Reliability data to use are given below:

Parameter	Value	Explanation (all cost amounts are given in NOK)
$MTTF_A$	8 000	Mean time to failure of pump A (hours) if we do not perform preventive maintenance.
q_B	0.1	The probability that pump B will not start when needed. In the first part of the exam paper, this value should be used. In subsequent tasks, the value should be calculated.
$MTTF_B$	16000	Mean time to failure of pump B (hours) if we do not perform preventive maintenance (overhaul). The MTTF value applies to the time the pump is in cold stand-by. We disregard the posibility that the pump may fail during operation, as the time it is running is very short.
α	3	Aging parameter of the pumps. Same aging parameter for both pumps. The time-to-failure of both pumps are assumed to be Weibull distributed.
c_{PM}	4 000	The cost of performing preventive overhaul task. The same value for both pumps. We assume that the preventive overhaul task means that we can consider the pumps almost as good as new after the work has been done.
c_{CM}	9 000	The cost of repairing a pump that has failed. This cost is greater than the preventive activity because the task cannot be planned and there may be consequential damage (the entire pump must be replaced).
c_{FT}	1 000	Cost of performing a function test (proof test) of pump B.
MDT	8	Mean downtime (hours) in case of pump A failure.
c_U	25 000	Production loss per hour the system does not produce.

We assume that the stand-by pump B cannot fail during operation since it will run for very few hours.

a) Write down the cost equation, i.e., the objective function, to determine the optimal maintenance interval of pump A assuming the fixed probability, q_B, that pump B will not start upon a demand.

b) Find an expression for the optimal interval and insert numeric values to find that interval, i.e., a numerical value of τ_A.

c) Explain how an optimal interval could be obtained by numerical minimization of the objective function, carry out such minimization and compare the result with the answer found in problem a).

4.8 We consider Problem 4.7 but will consider the maintenance of pump B. We will perform both an overhaul with interval $\tau_{B,O}$, and a proof-test with interval $\tau_{B,FT}$, where $\tau_{B,FT} < \tau_{B,O}$.

a) Give arguments for why we would use both of these maintenance activities, i.e., proof-test and overhaul. Hint: Time-to-failure of the stand-by pump during stand-by is assumed to be Weibull distributed.

b) Write down an expression for the effective failure rate of the stand-by pump as a function of $\tau_{B,O}$. Then use this expression as the "failure rate", i.e., "λ" in the unavailability formula: $q = \text{PFD} = \lambda\tau/2$. Insert numerical values, and calculate the probability that the stand-by pump will not start if $\tau_{B,O} = 14\,000$ and $\tau_{B,FT} = 7\,000$. Compare with q_B used in problem Problem 4.7.

c) Write down the cost function where you now treat the cost as function of all maintenance intervals, i.e., $C = C(\tau_A, \tau_{B,O}, \tau_{B,FT})$. Hint: Remember that you have to account for overhaul and proof-test of the stand-by pump, and the cost of repairing the stand-by pump.

c) Find a simultaneous optimal solution for all the maintenance activities, i.e., minimize the cost function wrt $\tau_A, \tau_{B,O}$ and $\tau_{B,FT}$. If you are not able to minimize with respect to $\tau_A, \tau_{B,O}$ and $\tau_{B,FT}$, keep τ_A equal to the value you found in Problem 4.7, keep $\tau_{B,O} = 14\,000$, and then minimize with respect to $\tau_{B,FT}$. Then keep keep τ_A equal to the value you found in Problem 4.7 keep $\tau_{B,FT}$ equal to the value you just found, and then minimize with respect to $\tau_{B,O}$. Finally keep $\tau_{B,O}$ and $\tau_{B,FT}$ equal to the values you just obtained, and minimize wrt τ_{BA}.

Chapter 5

Predictive Maintenance

5.1 Introduction

In contrast to traditional calendar based preventive maintenance the main idea of a predictive maintenance strategy is to utilize component condition, future loads, and opportunity windows to determine a "just in time" plan for maintenance. Condition information is basically used for:

- Anomaly detection, i.e., early warning of coming events
- Diagnostics, i.e., the search for root causes behind symptoms observed
- Prognostics, i.e., estimation of degradation rate, time to failure, remaining useful life etc. based on relevant information

A key element of PdM is use of sensor technology to assess the condition of the equipment. Condition monitoring is a method to evaluate the condition of equipment by performing periodic (offline) or continuous (online) equipment monitoring. In addition to the condition monitoring system (CMS) predictive maintenance often involves data from the supervisory control and data acquisition systems (SCADA). Data from the computerized maintenance management system (CMMS) is crucial in order to make predictions, i.e., the prognostics part.

In this presentation main focus is on the prognostics part of predictive maintenance.

5.2 The PF-model

The PF-model is one of the classical models within predictive maintenance. The basic idea in our context is that failure is regarded as a two- stage process. First, at some time a defect in the system becomes detectable, i.e., a potential failure (P), then, after some delay-time, the system fails due to the degeneration of the defect, i.e., a failure F. Backer and Christer (1994) present an exhaustive review of the models based on the the PF interval concept.

Figure 5.1 illustrates the situation behind the PF-model. On the y-axis we use the term 'health indicator' whereas in other presentations the term 'failure progression' is used. The idea is that there is "something" that could be used to spot a coming failure.

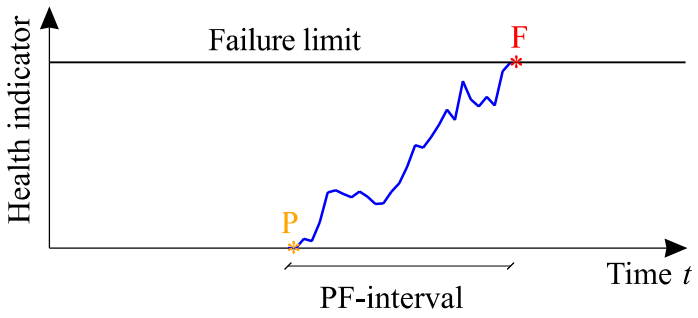


Figure 5.1: PF-Model

The point "P" depicts a potential failure, i.e., the time where a coming failure is observable. The time interval from the failure is first observable, and till a failure occurs is very often denoted the PF-interval. We will in the following denote this situation the "PF" situation because the PF-interval will be central in the understanding of effective maintenance strategies. An example could be a rail which is exposed to a combination of fatigue and a flat wheel which initiates a crack (potential failure, P). However, such cracks could be detected by ultrasonic inspection, and hopefully we will detect the crack before it propagates to a rail breakage, i.e., a failure (F). Note that if no maintenance is carried out, the time to failure will have an increasing failure rate (IFR).

To establish the effective failure rate we recognize that the two point of times "P" and "F" in Figure 5.1 are stochastic variables. This means that it is random when a potential failure occurs and the time it takes before it develops to a failure. The PF-interval is therefore also stochastic, and is denoted T_{PF}. As an example consider a rail where a crack can be initialized at different places of the rail, and thus time before the crack "reaches the surface" will vary. Another situation is where the crack propagation depends on the load, e.g., the number of heavy axels passing the track.

Periodic inspection is conducted at intervals of length τ to detect potential failures. The length of the inspection intervals should not be longer than the average PF-interval. However, since the PF-interval varies from time to time, and because there is also probability that a potential failure is not revealed during an inspection, the inspection interval should be shorter than the average PF-interval. A prerequisite for using the PF-intervals in maintenance planning is that a failure is alerted by some degradation in performance, or some indicator variable is alerting about the failure. Such a variable could be vibration, cracks, increased temperature etc.

The following quantities will be relevant when calculating the effective failure rate as a function of the maintenance interval:

- Mean PF-interval length, $\mathrm{E_{PF}}$
- Standard deviation in PF-interval length, $\mathrm{SD_{PF}}$
- Probability that an existing crack (or another warning situation) will be detected by an inspection, $p_\mathrm{I} = 1 - q_\mathrm{I}$, given that it is possible to detect the crack by condition monitoring method
- Coverage of the inspection method, i.e., percentage of cracks that could be detected, PC
- Interval length between inspections, τ
- Frequency of potential failures, f_P

In appendix E an expression for $Q_0(\tau, \mathrm{E_{PF}}, \mathrm{SD_{PF}}, q_\mathrm{I})$ is derived. This function represents the probability that the maintenance strategy fails to reveal a potential failure in due time. It is required to program this function in Excel VBA, Python or another programming language. The effective failure rate can now be calculated by:

$$\lambda_\mathrm{E}(\tau) = f_\mathrm{P} Q_0(\tau, \mathrm{E_{PF}}, \mathrm{SD_{PF}}, q_\mathrm{I}) \tag{5.1}$$

In order to obtain the optimal inspection interval we also need the rate of renewals:

$$\rho_\mathrm{E}(\tau) = f_\mathrm{P} \left[1 - Q_0(\tau, \mathrm{E_{PF}}, \mathrm{SD_{PF}}, q_\mathrm{I})\right] \tag{5.2}$$

The cost function to minimize is given by:

$$C(\tau) = C_\mathrm{I}/\tau + \lambda_\mathrm{E}(\tau) C_\mathrm{F} + \rho_\mathrm{E}(\tau) C_\mathrm{R} \tag{5.3}$$

where C_I is the cost of inspection, C_F is the total cost of a failure, and C_R is the cost of renewal, i.e., the cost of fixing a potential failure before it has developed to a failure.

Example 5.1 Ultrasonic inspection of rails

Rail breakages are a serious threat to railway safety, and periodic ultrasonic inspection of the rails is a required safety barrier for safe operation of the railway infrastructure. The objective of inspection of the rails are to detect presence of defects such as cracks and track misalignments.

Cracks are initiated within the rail and, as the rail operation proceeds, they worsen if no recovery action is undertaken. An ultrasonic inspection car is used to detect potential defects. Candidate defects are verified by a manual inspection with a hand-held trolley with ultrasonic

inspection equipment. Defects, or cracks, are assigned a "severity class" and a corresponding maintenance procedure is currently undertaken:

2b Keep rail under observation, and perform a new inspection every 3 MBT

2a Keep rail under observation, and perform a new inspection every 1 MBT

1 Repair the defect quickly, i.e., within one month

0 Repair failure immediately and initiate traffic restrictions until failure is fixed.

where MBT is million gross tonnage passed at the specific location. A zero-defect is considered to be a failure, and would develop to a rail breakage in short time, i.e., a state F. From an optimization point of view, both the frequency of inspection and the follow up regime should be optimized.

In this example we will only consider the frequency of running the inspection car. See Vatn (2023) for a more comprehensive study also considering follow-up strategies for the various defect states.

We consider a railway line where the rails are approaching the technical life time. With technical life time we mean that the rails are worn-out, and the rate of fatigue cracks are increasing.

The rate of new cracks detected by the ultrasonic rate (the rate of potential failures) is currently **0.5** failures per **10** km per year. The number of potential defects that cause rail breaks depends on the inspection interval. Of the rail breakages, we again assume that **5%** gives derailment. The cost of a derailment is on average NOK **15 million**. Correction of potential defects (before they result in a rail breakage) costs NOK **20,000**, while repair after rail breakage costs NOK **40,000**. The expected PF interval length is assumed to be **5** years , and the standard deviation of the PF interval is assumed to be **3** years. It costs NOK **4,000** per 10 km of ultrasound inspection that can reveal potential errors. The probability that a defect (potential failure) is not detected by the ultrasonic train is **20%** per run.

The $Q_0(\tau, \mathrm{E_{PF}}, \mathrm{SD_{PF}}, q_\mathrm{I}$ function is implemented in MaintOp.xlsm , and Table 5.1 shows the parameter used.

Figure 5.2 shows the cost contributions per 10 km for the various cost elements. The optimal inspection interval is $\tau^* = 1.2$ years if found by using the Solver.

5.3 Predictive maintenance and Cox-proportional models

Predictive maintenance is about utilizing information regarding the *condition* of a component and the *future expected loads* in order to judge the correct time for intervention. In the previous section a simple model was derived but the current condition of the component and the future expected loads were not explicitly used. Some formalism is required for such a utilization.

Table 5.1: Parameter values specified in Excel

Parameter	Value	Formula/Value
c_R	20000	20000
c_CM	40000	40000
c_Safety	750000	=0.05*15000000
c_I	4000	4000
c_F	790000	=C_CM+C_Safety
f	0.05	0.05
e_pf	5	5
sd_pf	3	3
q	0.2	0.2
tau	1.2	1.2
Q_0	4.278E-02	=Q0(tau,e_pf,sd_pf,q)
C(tau)	5910	=c_I/tau + f*Q_0*c_F+f*(1-Q_0)*c_R

This will be crucial for digital twins where a computerized mathematical model of the system is established where real time information regarding state, production profile and plans etc are connected via internet of thins (IoT).

A reasonable simple extension of the model used in the previous section will be derived. The starting point is the failure rate function, $z(t)$. We stick to the Weibull distribution where the failure rate function is given by $z(t) = \alpha\lambda^{\alpha} t^{\alpha-1}$. We observe that $z(t)$ does not contain neither the current state nor the future loads. The so-called Cox-proportional hazard model Cox (1972) is often used to incorporate the current state in the failure rate function.

It should be noted that in a Cox-proportional hazard model we will utilize the current state measured by some health indicator and future loads in the model for time-to-failure. However

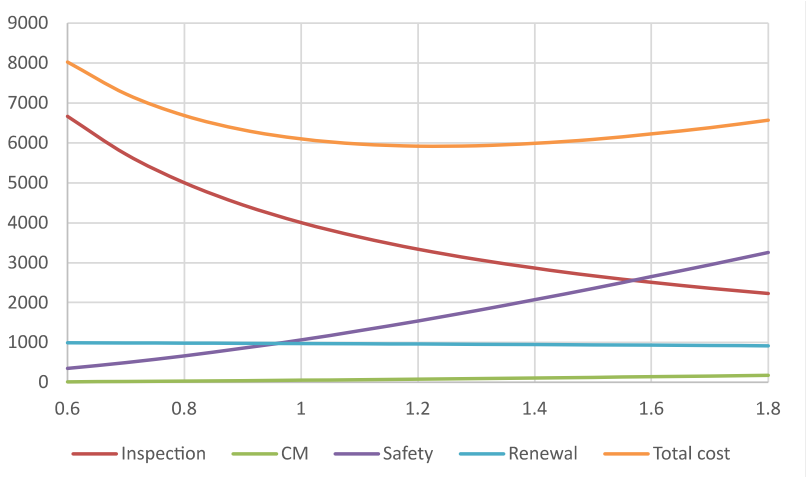


Figure 5.2: PF-Model

in this model we do not explicit model how the health indicator will develop from now on. This means that there is a "here and now" assessment of the time-to-failure distribution. Typically such a model could be established from statistical data. Later on we will introduce the Wiener and gamma processes where we explicit make a model for how the health indicator will develop.

Let $\mathbf{y}$ be the vector of current relevant state information for the component, for example temperature, vibration level and so on and let $\overline{\mathbf{x}(t)}$ be the vector of average loads in the time period $[0, t)$. The failure rate function may be written on the form:

$$z(t|\mathbf{y}, \overline{\mathbf{x}(t)}) = z_0(t) e^{\boldsymbol{\beta}_1 \mathbf{y}} e^{\boldsymbol{\beta}_2 \overline{\mathbf{x}(t)}} \tag{5.4}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are regression coefficient vectors established by for example statistical analysis of data. $z_0(t)$ is a baseline failure rate function, typically on the form $z_0(t) = \alpha\lambda^\alpha t^{\alpha-1}$. t is running time measured from the current time, say t_0.

Now assume that the parameters α, λ, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are all known. Further assume that the current component state, $\mathbf{y}$, is known and that we have an estimate of future load $\overline{\mathbf{x}(t)}$. The cost equation to minimize is:

$$C(t) = c_{\mathrm{PM},0} e^{-t/\theta} + c_{\mathrm{U}} F_T(t|\mathbf{y}, \overline{\mathbf{x}(t)}) \tag{5.5}$$

where the cumulative distribution function is given by:

$$F_T(t|\mathbf{y}, \overline{\mathbf{x}(t)}) = 1 - \exp\left(- \int_0^t z(u|\mathbf{y}, \overline{\mathbf{x}(u)})\, du \right) \tag{5.6}$$

A main objective when establishing digital twins for maintenance and operations is to set up a regime for data collection and analysis. It is beyond the scope of this presentation to describe relevant statistical methods. Typically a partial likelihood estimation approach is recommended where the impact of the regression coefficient is estimated, and then a separate approach is used for estimation of the failure rate function. See e.g., Cox (1972).

If no data is available we might use expert judgements for elicitation of the relevant model parameters. As a basis for our argument we will use the PF-model illustrated in Figure 5.1 as a conceptual model. The history up to the potential failure is now of limited value, the only is the current state and future loads. We assume that the potential failure as just occurred. Let T be the length of the PF-interval, and let $F_T(t|\mathbf{y}, \overline{\mathbf{x}(t)})$ denote the cumulative distribution function of the PF-interval. If we do not have statistical data to estimate the model parameters, the following procedure may be used for elicitation of the model parameters:

1. Assess the expected length of the PF-interval under the assumption of insignificant future load $\overline{\mathbf{x}(t)}$. Denote this value by ξ.

2. Asses the consistency of the PF-interval by the shape parameter α in the Weibull distribution. As a rule of thumb use
   - $\alpha = 2$ corresponds to a variety of failure mechanisms and causes leading to a failure.
   - $\alpha = 3$ corresponds to a few failure mechanisms and causes leading to a failure.
   - $\alpha = 4$ corresponds to a rather specific failure mechanism / failure cause leading to a failure.
3. Calculate the intensity parameter by $\lambda = \Gamma\left(1/\alpha + 1\right)/\xi$.
4. For each y_i in $\mathbf{y}$ let $y_i = 0$ correspond to the condition at the point of time P in Figure 5.1. This corresponds to no significant damage or degradation for the actual regression variable.
5. For each y_i in $\mathbf{y}$ let $y_{i,\mathrm{C}}$ be a critical value for that particular regression variable. Under the assumption that all other regression variables $y_j = 0, j \neq i$ assess the reduction in ξ by some factor, say k_i. Note that there is no specific "rule" to determine $y_{i,\mathrm{C}}$, and the higher value chosen, the lower value will be assessed for ξ.
6. Calculate the corresponding regression parameters by $\beta_i = -(\ln k_i)/y_{i,\mathrm{C}}$, i.e., for the elements in $\boldsymbol{\beta}_1$.
7. Repeat the procedure for each $\overline{x_i(t)}$ in $\overline{\mathbf{x}(t)}$ to find the elements of $\boldsymbol{\beta}_2$.

Figure 5.4 illustrates the idea behind the factor k_i. The expected length of the PF-interval equals ξ if the state variable $y_i = 0$. Then we can imagine a situation where $y_i = y_{i,\mathrm{C}}$. With such a critical value of the state variable y_i the expected PF-interval is much shorter, say $k_i\xi$. From such an argument we can determine the factor k_i.

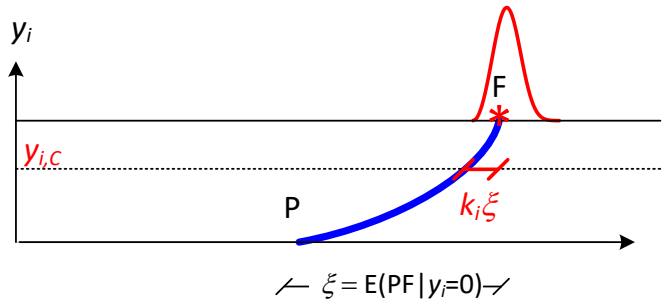


Figure 5.3: Elicitating of the factor k_i

Example 5.2 Using vibration data and expected average loads

If the PF-model in Figure 5.1 is found realistic for the maintenance challenge at hand, there are

two challenges. The first one is to determine the inspection strategy to reveal potential failures. The second challenge is the response to a revealed potential failure. Often it is not possible to fix a potential failure immediately, or at least it will be very costly. Fixing the problem requires planning, access to spare parts, resources etc. To model this, we assume that there is an upper limit for repair/renewal cost, say $c_{R,0}$. This cost represents the cost if the repair/replacement cost is carried out more or less immediately after the potential failure has been revealed. Then we assume that the cost will become lower if we could wait t time units before we do the work. Various models for the drop in cost could be used, but in the following we assume an exponential drop. Further if a failure occurs before t we have to pay some unplanned failure cost, say c_U. This means that if we decide to wait t time units from now on until we will execute the work, the expected cost as a function of t is given by:

$$C(t) = c_{R,0}e^{-t/\theta} + c_U F_T(t) \tag{5.7}$$

In Equation (5.7) neither the current condition nor the future load is considered. Let y be the vibration level measured by the so-called “RMS" value (Root Mean Square) which is an ISO convention. Technically the RMS value is calculated by multiplying the peak amplitude by 0.707. For machines of medium size the vibration level is mapped into zones where zone A is the normal level which we here assume corresponds to $y = 0$, zone B which still is considered acceptable ranges from $y = 1.8$ to $y = 4.5$, zone C which is critical ranges from $y = 4.5$ to $y = 11.2$ and zone D corresponds to $y > 11.2$. A machine in zone D is considered to have serious damages within very short time and is therefore often protected by a protection system causing the machine to shut down (TRIP).

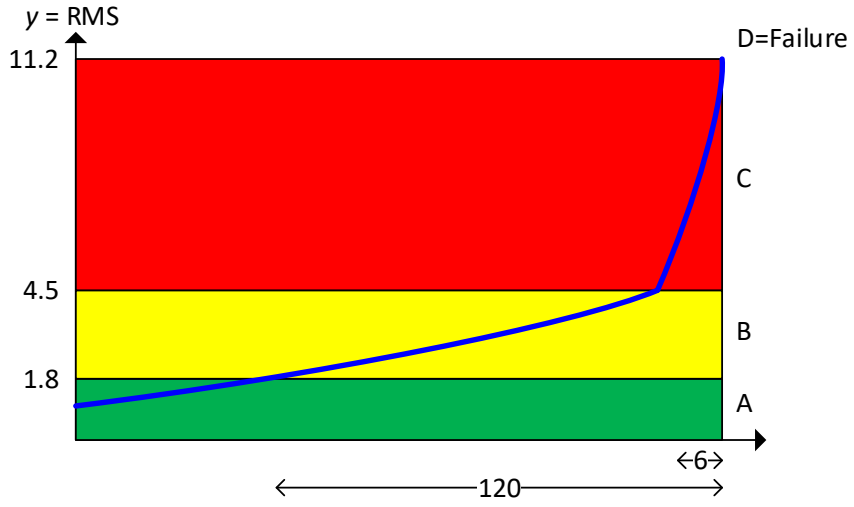


Figure 5.4: Levels for the root mean square based on ISO

The future operational load is specified by the variable x which measures the portion of time the machine is run on more than 90% of maximum capacity. A high value of x is expected to increase the degradation rate and hence give a shorter value of the PF-interval.

In the elicitation process the maintenance engineer assess the mean residual time to failure, i.e., the time until the protection system will trip the system (PF-interval) to be $\xi = 120$ days when an anomaly situation occur, i.e., drifting into zone B. Since only vibration and excessive load is considered as influencing factors of the PF-interval the shape parameter is assessed by $\alpha = 4$. This gives $\lambda = \Gamma(1/\alpha + 1)/\xi = \Gamma(1.25)/120 \approx 0.00775$.

For the elicitation of the k-factor for the RMS a critical value for the vibration is set to $y_C = 4.5$. The expert now assess that given such a value, the expected number of days until a trip occurs is 6 days. The corresponding reduction factor for the remaining time to trip is assessed to $k_Y = 6/120 = 0.05$. This gives $\beta_Y = -(\ln k_Y)/y_C = -(\ln 0.05)/(4.5 - 1.8) \approx 0.05$. Note that we use 4.5-1.8 rather than 4.5 because 1.8 is considered as the "zero-point" for y corresponding to level A is considered to be normal vibration level.

A machine running with 90% of maximum capacity or more in $x_C = 0.25 = 25\%$ of the time is assessed to have a reduction factor of $k_X = 0.1$ (12 days to failure in average). This gives $\beta_X = -(\ln k_X)/x_C = -(\ln 0.1)/.25 \approx 9.2$.

The relevant parameters required to calculate the cumulative distribution function for the PF-interval in Equation (5.6) have now been established. Now assume that we have observed $y = 4$ and from the production it is desired to run on high load in 10% of the time, i.e., $x = 0.1$.

The cumulative distribution function in Equation 5.7 did not include the explanatory variables. Further there is no extra profit related to running at "full speed", i.e., running at a higher load than 90%. To access the extra profit we assume that if the machine is run at 90% load or more, then the extra profit that day is p_{HL} Assuming a linear relation the total extra profit by running at high load in x portion of time for t days is $p_{HL}xt$. This amount is then subtracted from the cost function, and we get:

$$C(t,x) = c_{R,0}e^{-t/\theta} + c_U F_T(t|y,x) - p_{HL}xt \tag{5.8}$$

where

$$F_T(t|x,y) = 1 - \exp\left(-\int_0^t z(u|y,x)\,du\right)$$

and

$$z(t|y,x) = \alpha\lambda^\alpha t^{\alpha-1}e^{\beta_Y(y-1.8)+\beta_X x}$$

Note that we use $y-1.8$ because the "zero-point" for y is 1.8 corresponding to level A is "normal vibration".

The cost/profit figures are as follows: $c_{R,0} = 15,000$, $c_U = 30,000$ and $p_{HL} = 1,000$ where all values are given in NOKs. The characteristic time in the decay function is given by $\theta = 30$ days. In the example we also assume that there is a maintenance window only once a week, and the first opportunity will be in 3 days.

Inserting in Equation (5.6) and using the cost function in Equation (5.5) Table 5.2 indicates that we should use the opportunity that comes after 31 days.

Table 5.2: Results for the Cox proportional hazard rate model

t	Repair	Downtime	Production	Total
3	13573	0	-300	13273
10	10748	33	-1000	9781
17	8511	273	-1700	7085
24	6740	1074	-2400	5413
31	5337	2907	-3100	5145
38	4227	6224	-3800	6650
45	3347	11185	-4500	10032

Note that the cumulative distribution function calculated by Equation (5.6) is the unconditional distribution function given we were at point of time P in Figure 5.1. In reality since $y = 4$ it is reasonable to believe that some days has elapsed since the potential failure was evident. A conditional distribution function is therefore more appropriate. This means that we also need to assess the time since the potential failure occurred. Let t_0 be the current time, and assume that the time since the potential failure (P) is s time units. Let t denote the running time from now on, i.e., t_0 corresponds to $t = 0$. Using the rule for conditional probabilities we obtain the following modified cost function:

$$C(t) = c_{PM,0}e^{-t/\theta} + c_U\left[1 - \frac{1 - F_T(t+s|\mathbf{y}, \overline{\mathbf{x}(t+s)})}{1 - F_T(s|\mathbf{y}, \overline{\mathbf{x}(s)})}\right] - p_{HL}xt$$

In the example calculation this conditional approach is not used. □

Example 5.3 Towards a real time model - The digital twin

The previous example is now used as motivation for developing a simple stochastic digital twin. A digital twin may be viewed as a digital simulation model with built in analytics, decision support, and self learning features. Learning features will not be discussed in this example, and only glimpse of analytics is provided.

The digital twin is represented by two models, one maintenance model and one production

model, where these models interact via the Internet of Things. In the following the maintenance model is denoted the maintenance twin and the production model is denoted the production twin. The physical counterpart of the maintenance twin is the actual component state, the physical load on the machine, the actual maintenance carried out the actual time the machine can not produce due to preventive and/or corrective maintenance and so on. The physical counterpart of the production twin is what is actually being produced, when the production takes place, the economic value of the production, the cost of production, the various machines being used, the use of personnel and resources and so on.

Let $\mathcal{T}$ be the operational windows for execution of a preventive maintenance task of the packing machine, i.e., the point of times $\tau_1, \tau_1 + \tau, \tau_1 + 2\tau, \ldots$. The decision support to be provided by the maintenance twin upon a potential failure situation is now:

$$\min_{t \in \mathcal{T}} C(t) = c_{\mathrm{PM},0} e^{-t/\theta} + c_{\mathrm{U}} F_T(t|\mathbf{y}, \overline{\mathbf{x}(t)} \tag{5.9}$$

The maintenance twin represented by Equation 5.9 has to be implemented on a digital platform, for example MS Excel. The maintenance twin needs to be fed with data from the production twin. Here the production twin is very simple, only a set of predefined scenarios combining different values of $c_{\mathrm{PM},0}, c_{\mathrm{U}}, \mathbf{y}$, and $\overline{\mathbf{x}(t)}$. Table 5.3 shows the data used in this simple MS Excel representation of the two twins interacting. In a real life implementation the data in Table 5.3 needs to be generated by the ERP system, the SCADA system and so on.

Table 5.3: Data used in the production twin

$c_{\mathrm{PM},0}$	c_{U}	y	x	CPS message/Comment
15000	35000	4	0.1	Base line (from example)
15000	35000	3	0.3	High future loads
5000	35000	3	0.1	Cheap PM due to low production
15000	35000	2	0.15	Lower degradation
15000	35000	4	0.15	Very high degradation
15000	100000	3	0.15	Very high failure cost

5.4 Gradual failure progression

The PF-model in Section 5.2 illustrates what we often denote a “fast failure progression” because the length of the PF-interval is relatively shorted compared to the time between potential failures. This means that there is no sign of degradation for a long time, and then some failure mechanism become evident and the item will fail within short time if nothing is done.

In this section we will introduce a general framework. The basis is still that maintenance is

condition based where the understanding of degradation of the item determines appropriate maintenance action and time for maintenance.

In the following we distinguish between:

- $\{X(t), t \geq 0\}$ is a stochastic process describing the actual degradation of the item at time t
- $\{Y(t), t \geq 0\}$ is a stochastic process describing the *measurements* of degradation of the item at time t

where the measurements typically contain noise. Degradation could be crack lengths, corrosion depths, vibration levels etc. In some situations it may be difficult to distinguish the variation in the degradation process from the measurement errors. In this presentation we will not explicitly consider imperfect measurements of the degradation in order to make simple presentations. We will therefore not make an explicit definition of what is the difference between $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$.

Figure 5.5 shows a typical example of the development into a failure. On the y-axis the figure shows $\{X(t), t \geq 0\}$ the actual degradation. As for the PF-model this would be some health indicator like crack length, corrosion level, wear etc. In addition to the failure limit we introduce a maintenance limit, i.e., when $\{X(t), t \geq 0\}$ exceeds the maintenance limit, a request for maintenance is put forward. An important decision variable is then what the maintenance limit should be.

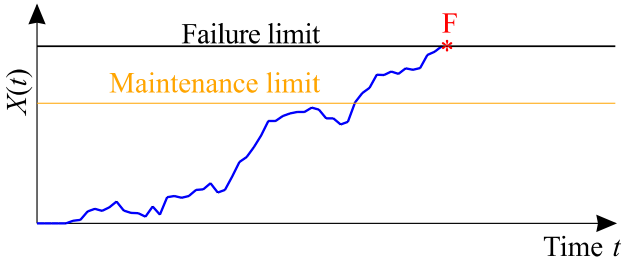


Figure 5.5: Gradual failure progression

5.5 Remaining Useful Lifetime

In degradation modelling (prognostics) the term Remaining Useful Lifetime (RUL) is introduced. RUL(t_0) is a stochastic variable that measures the time from t_0 until the item is not "useful" any more. The term 'Useful' needs to be defined, for example a failure, or some other bad performance. Since RUL is a stochastic variable, we often need the distribution function, i.e.,

$$\Pr(\text{RUL}(t_o) \leq t) = F_{\text{RUL}(t_0)}(t) \tag{5.10}$$

where t_0 is the current time, and t is a future point of time, typically measured from t_0 as the starting point.

A huge number of mathematical models for RUL prediction exist in the literature. In the following a limited number of ideas are pursued. The definition in Equation (5.10) will not help us since there is no explicit link to the condition or degradation of the item. A more explicit definition of RUL is therefore:

$$\text{RUL}(t_0) = \min\{h : X(t_0 + h) \in \mathscr{X}_l\} \tag{5.11}$$

where $\mathscr{X}_l$ is the set of states where the item is considered not useful. In Figure 5.5 this corresponds to all values above the "Failure limit". The RUL distribution (CDF) is defined as:

$$\Pr(\text{RUL}(t_0) \leq t) = \Pr(\min\{h : X(t_0 + h) \in \mathscr{X}_l\} \leq t | T > t_0, Y(\tau)_{\tau \in \mathscr{T}_{t_0}}) \tag{5.12}$$

where we condition on the fact that the item is still useful ($T > t_0$), and the knowledge of the measurements, i.e., the various observations (Y). In Equation (5.12) $X(t_0+h)$ is the *actual* degradation h time units ahead of current time t_0. When we condition on $Y(u)_{u \in \mathscr{T}_{t_0}}$ this means that the only information we have at the current time t_0 is the *measurements*, i.e., the process $\{Y(u)\}$ sampled at various points in time, i.e., the set $\mathscr{T}_{t_0}$. In principle we do not know the actual states of the system up to the current time, but since $T > t_0$ we for sure know that $X(u) \notin \mathscr{X}_l, u \leq t_0$.

The Wiener and gamma processes are popular stochastic processes used to model degradation. Both processes assume that the change in degradation level in a small time interval can be described by a stochastic variable. In the Wiener process these changes can be both positive and negative, whereas in the gamma process the changes are always positive, i.e., positive increments. There are various pros and cons for these two processes. The gamma process is more intuitive, since increments (degradation) is always positive which is true for man failure mechanism, i.e., we can not improve unless some measures are taken. On the other side, measurements of the degradation often show that the change in degradation level from one point of time to the next may be negative. This could then be caused by measurement errors (noise).

Both the Wiener and gamma processes assume that $\{X(t), t \geq 0\}$ can take any value in some range. In some situations we limit the value of the process to a countable number of values. The Markov process is such a process also very relevant for maintenance modelling.

5.6 Wiener Process with Linear Drift

Before we define a Wiener process with drift we define the Wiener process $\{W_t, t \geq 0\}$ by:

1. $W_0 = 0$

2. W has independent increments: for every $t > 0$, the future increments $W_{t+u} - W_t, u \geq 0$, are independent of the past values $W_s, s \leq t$.

3. W has Gaussian increments: $W_{t+u} - W_t$ is normally distributed with mean 0 and variance $u, W_{t+u} - W_t \sim \mathcal{N}(0, u)$.

4. W has continuous paths: W_t is continuous in t.

Note the slightly different notation where in some situations we use $\{W_t, t \geq 0\}$ and in other situations we use $\{W(t), t \geq 0\}$.

We now define stochastic process:

$$X_t = \mu t + \sigma W_t$$

as a Wiener process with linear drift μ and infinitesimal variance σ_2.

It follows that $X_t = X(t)$ is normally distributed with mean μt and variance $\sigma^2 t$. Further X has Gaussian increments: $X_{t+u} - X_t$ is normally distributed with mean μu and variance $\sigma^2 u$, i.e., $X_{t+u} - X_t \sim \mathcal{N}(\mu u, \sigma^2 u)$.

It is well known from the theory of stochastic processes that the time T when the process for the first time reach the level ℓ is inverse-Gauss distributed with parameters $\alpha = \ell/\mu$ and $\beta = (\ell/\sigma)^2$.

For the inverse-Gauss distribution, i.e., $X \sim \text{IG}(\alpha, \beta)$ we have:

$$f_X(x; \alpha, \beta) = \sqrt{\frac{\beta}{2\pi x^3}} \exp\left(-\frac{\beta(x-\alpha)^2}{2\alpha^2 x}\right) \tag{5.13}$$

and

$$F_X(x; \alpha, \beta) = \Phi\left(\frac{\sqrt{\beta}}{\alpha}\sqrt{x} - \sqrt{\beta}\frac{1}{\sqrt{x}}\right) + \Phi\left(-\frac{\sqrt{\beta}}{\alpha}\sqrt{x} - \sqrt{\beta}\frac{1}{\sqrt{x}}\right) e^{2\beta/\alpha} \tag{5.14}$$

The expected value and variance are given by:

$$\begin{aligned} \text{E}[X] &= \alpha \\ \text{Var}(X) &= \alpha^3/\beta \end{aligned}$$

In the Wiener process with parameters μ, σ the time, T to first passage of the threshold ℓ is then

$$T \sim \text{IG}(\ell/\mu, (\ell/\sigma)^2)$$

and the expected value and variance are given by:

$$\mathrm{E}[T] = \ell/\mu$$
$$\mathrm{Var}(T) = \sigma^2 \ell/\mu^3$$

5.6.1 Maintenance decision problem

Figure 5.6 shows the Wiener process with drift and gives the motivation for the maintenance decision problem. The elements in the model is discussed in the following.

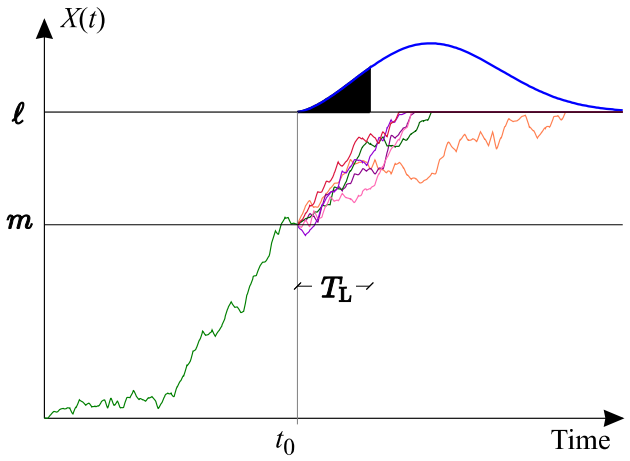


Figure 5.6: Maintenance model with deterministic lead time T_{L}

We consider the following situation:

- The degradation process can be monitored continuously without any uncertainty, and the degradation level at time t is $X(t)$
- A failure occurs the first time $X(t) \geq \ell$
- When degradation reaches the maintenance limit, m a request is placed to replace the component with a new component, in Figure 5.6 the maintenance limit is reached at time t_0
- There is a deterministic lead time, say T_{L}, i.e., the time elapsed from the replacement request is placed until it is executed
- The objective is to determine the maintenance limit, $m < \ell$, i.e., how close to the failure limit we dear to go

- Figure 5.6 shows one trajectory of the degradation up to the maintenance limit, and from three on several trajectories are shown to illustrate the randomness in the degradation. If $X(t) \geq \ell$ for some $t < t_0 + T_L$ the item fails before it is replaced which is indicated with the filled area under the RUL-distribution (Remaining Useful Life).

The cost equation to minimize is:

$$C(m) = \frac{c_R + c_F F(T_L|m) + c_D \int_0^{T_L} f(t|m)(T_L - t)\,dt}{\text{MTBR}(m)} \tag{5.15}$$

where

- c_R = cost of renewal/replacement
- c_F cost of failure (additional cost for corrective maintenance and extra cost for the failure event)
- c_D = cost per hour down time
- $F(t)$ and $f(t)$ are CDF and PDF for the remaining useful lifetime (RUL), given we are at the maintenance limit m at some point
- MTBR(m) = Mean Time Between Renewals, given the decision rule to request a maintenance at m

We now consider one maintenance cycle:

- Assume that we at time t in this cycle observe $Y(t) = m$
- Let RUL_m be the time from t until a failure occurs
- RUL_m is inverse-Gauss distributed with parameters $\alpha_m = (\ell - m)/\mu$ and $\beta_m = (\ell - m)^2/\sigma^2$, where μ and σ^2 are the parameters in the Wiener process, and ℓ is the failure threshold
- Thus, $F() = F(t; \alpha_m; \beta_m)$ = and $f() = f(t; \alpha_m; \beta_m)$ are given by equations (5.14) and (5.13) respectively, and the nominator of $C(m)$ may be obtained by numerical integration
- $\text{MTBR}(m) = m/\mu + T_L$

5.6.2 Operational load and relaxing on the use of the item

When we reach the maintenance limit we could relax on production, e.g.,:

- Produce less items

- Stop a wind turbine when wind speed > $15 m/s$

Let x be a measure of how much we relax on production, i.e., a decision variable, and let $c_{\mathrm{Rx}}(x)$ be corresponding production loss per unit time. The cost equation to minimize now is:

$$C(m,x) = \frac{c_{\mathrm{R}} + c_{\mathrm{F}} F(T_{\mathrm{L}}|m,x) + c_{\mathrm{D}} \int_0^{T_{\mathrm{L}}} f(t|m,x)(T_{\mathrm{L}} - t)\,dt + c_{\mathrm{Rx}}(x)\,T_{\mathrm{L}}}{\mathrm{MTBR}(m_{\mathrm{L}})} \tag{5.16}$$

where $F()$ and $f()$ now depends on x as well as the maintenance limit.

Two aspects need to be considered:

- The impact of the relax on the degradation rate
- The impact on direct profit

The relax "decision variable" is denoted x, and we have assumed

- $\mu(x) = \mu_0(1-x)$
- $\sigma(x) = \sigma_0(1-x)$
- $C_{\mathrm{Rx}}(x) = 0.002 c_{\mathrm{U}}(x + 25x^2)$

In the example, we "relaxed" on our selves, i.e., the turbine which is approaching a fault state. In light of "wake effects", it might be more relevant to consider relaxing on the "front runners". Two aspects need still to be considered:

- The impact of the relax on the degradation rate
- The impact on direct profit

The starting point for the wake model is the classical engineering model by Jensen (1983) who proposed a model for reduced wind speed downstream

$$U_J/U_0 = 1 - 2a\left(R/(R + \alpha_J X)\right)^2 \tag{5.17}$$

where R = rotor diameter, and X = distance between two turbines. Often $a = 1/3$, and α_J=0.05 for offshore wind is used. This gives

$$U_J/U_0 = 1 - \frac{2}{3}\left(\frac{R}{R + 0.05X}\right)^2 = 1 - \frac{2}{3}\left(\frac{1}{1 + 0.05X/R}\right)^2 \tag{5.18}$$

For normal spacing of the turbines, the reduction factor U_J/U_0 could be between 0.7 and 0.8.

First of all we should acknowledge that engineering models are not very accurate, and in particular to consider the situation at the wind farm level, such a model might be too simple. To optimize production, independent of the impact of degradation, yawing could impact the loss in inn speed. We do not propose models here, but yawing means reduced swiping area of the frontrunner turbine, and a reduction factor for the front runner could be something like $\cos\phi$, where ϕ is the yawing angle. However, the total impact on the effect this will have on the downstream wake profile is not that easy to model, and far beyond the scope of this presentation

From the maintenance perspective, the turbulence is our main concern wrt the wake effect. In the wake shadow, it is expected to be much turbulence. We could may be use something like the inverse reduction factor, i.e., U_0/U_J as a starting point for an "increase" factor of e.g., fatigue loads.

Note the difference:

- General increased load due to turbulence, and how we consider this as a part of the overall objective function for wind farm control
- The explicit modelling of a given situation, where we have observed a critical degradation, and the aim is to reduce the risk of failure until maintenance could be carried out, i.e., our example

5.7 Gamma process

A stationary *gamma* process $X(t), t \geq 0$ is defined by:

1. $X(0) = 0$
2. $X(t), t \geq 0$ has independent and stationary increments
3. The increments in an interval $(s, t]$ is $X(t) - X(s)$ and are assumed to be gamma distributed with parameters $(t-s)\alpha$ and β. α and β are denoted the shape and scale parameters respectively.

For the gamma process it is straight forward to obtain the cumulative distribution function for the first hitting time. Assume degradation degradation of an item follows a stationary *gamma* process $X(t), t \geq 0$ and the item will fail the first time $X(t)$ exceeds a failure threshold ℓ. The cumulative distribution function for the time-to-failure is:

$$F(t) = \Pr(T \leq t) = \Pr(X(t) \geq \ell) = 1 - G(\ell; \alpha t, \beta) \tag{5.19}$$

where $G(x; a, b)$ is the cumulative distribution function for the gamma distribution with parameters a and b. The results follows from the fact that the increments are gamma distributed

The expected degradation in a time interval of length $s-t$ is $(t-s)\alpha/\beta$. Referring to the example given for the Weibull process, we could be tempted to assume that the expected time for a new item to reach the maintenance threshold, ℓ is $\ell\beta/\alpha$. This is not the case and it can be shown that the expected value can be approximated by $\ell\beta/\alpha + 1/(2\alpha)$. The extra term $1/(2\alpha)$ is often denoted "overshooting" effect. The idea is that the gamma process is a jump process. This means that it will never exactly hit the value ℓ but rather hit slightly above, and hence it takes some "extra" time compared to if it was an "exact hit".

Compared to the Wiener process, it is however easier to find the RUL_m distribution. Assume we order a maintenance when the process reaches the value m. We then have

$$F_{\text{RUL}_m}(t) = \Pr(\text{RUL}_m \le t) = \Pr(X(T_m + t) \ge \ell | X(T_m) < \ell, \text{ history up to } T_m) \tag{5.20}$$

$$\approx \Pr(X(T_m + t) - X(T_m) \ge \ell - m) = \tag{5.21}$$

$$\int_{\ell-m}^{\infty} f_{\alpha t,\beta}(u)\,du = 1 - F_{\alpha t,\beta}(\ell - m) \tag{5.22}$$

where T_m is the point of time when the process exceeds the maintenance limit. $f_{\alpha t,\beta}(\cdot)$ and $F_{\alpha t,\beta}(\cdot)$ are the PDF and CDF of the gamma distribution with parameters αt and β respectively.

Note the approximation which is due to the fact that we never exactly reach the maintenance limit m due to overshooting. If we pursue the maintenance model used in the Wiener process example, we should also take the "overshooting" into account for the expected time to reach the maintenance limit which could be approximated by $\text{MTBR} = m\beta/\alpha + 1/(2\alpha) + T_{\text{L}}$. In the cost model we also need the PDF for the RUL in addition to the CDF derived above.

5.7.1 Response Time

To demonstrate the use of the Gamma process we present a situation similar to the one in Section 5.6. As a starting point we assume continuous condition monitoring where the inspection interval is not relevant for optimization. The challenge is to look at an optimal response time after a critical situation has been identified after analysing the data from the continuous condition monitoring. The model developed below can also be used in conjunction with the optimization of inspection intervals, but this is not pursued here. We make the following assumptions:

- The degradation process of a component or a system follows a gamma process with parameters α and β and a failure threshold at ℓ
- At the time t_0 a critical development of the condition is observed. t_0 can be the current time, or some anticipated time in the future.
- We assume that in a rather short time, i.e., at time $t_0 + x_{\text{O}}$, there will be a reasonable good

opportunity to execute an improvement activity that brings the system back to a significant improved condition

- As a simplification, we assume that x_O is known
- In some situations, there is no natural "opportunity" to carry out the improvement, and it may be natural to set a fictitious value for x_O, typically a value where cost of the improvement activity does not decrease even if we wait longer than x_O to carry out the improvement
- We also assume that it is possible to execute the activity prior to $t_0 + x_O$ but at a higher cost since "priority" is usually not free of charge
- The cost to perform an improvement, $C_R(x)$ is assumed to dependent on *when* the improvement is carried out, i.e., x time units from t_0. The faster it needs to be implemented, the higher the expected cost. The lowest cost is assumed at the time $t_0 + x_O$.
- If the situation develops rapidly, we may experience e failure before the improvement activity is scheduled to be carried out. If this happens, we assume that the system is unavailable until the planned time, i.e. at the time $t_0 + x_O$. During this period, there is an unavailability cost, c_D per unit time. In addition, there is a fixed cost associated with the failure itself, c_F. c_F represents both additional costs associated with repairing a failure compared to the planned repair activity, as well as other costs associated with the failure itself.
- The longer you wait to fix the critical development of the condition, the more you save in average renewal costs. Let c_r denote average renewal costs per unit time.
- Let $f(t)$ denote the probability density function from t_0 until a failure occurs. $f(t)$ will typically depend on the current condition, although we do not notationally express this in $f(t)$.

Figure 5.7 illustrates the situation. At time t_0 degradation has reached a critical level. At time $t_0 + x_O$ there is an opportunity to carry out the activity at the lowest possible cost. The cost function, $C_R(x)$, drops from the highest value at $x = 0$, i.e., at time t_0 and drops from here on until time $t_0 + x_O$. Above the failure threshold the probability density function of the time-to-failure is indicated. Various random trajectory from the critical level to the failure threshold is indicated.

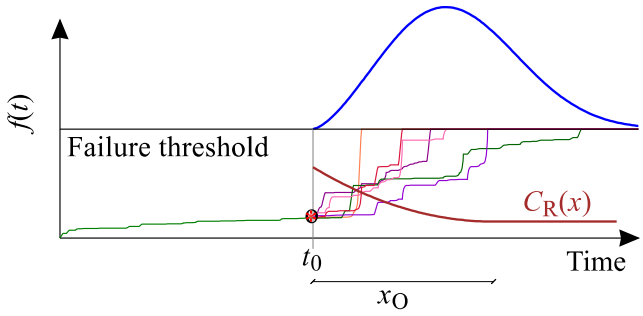


Figure 5.7: Optimizing response time x

The decision problem is to find the optimal time for the execution of the improvement activity, x, i.e., the optimal response time. A realistic cost function is given by:

$$C(x) = C_R(x) + (x_O - x)c_r + c_D \int_0^x f(u)(x-u)\,du + c_F \int_0^x f(u)\,du \tag{5.23}$$

where $f(u)$ is the probability density function for RUL, given the knowledge and condition at time t_0. Numerical methods are needed both to calculate $C(x)$, and to minimize wrt x.

For the last integral we have $\int_0^x f(u)\,du = F(x)$ where $F(\cdot)$ is the cumulative distribution function for RUL. In some cases it is easier to calculate the CDF than the PDF, hence the integral $\int_0^x f(u)(x-u)\,du$ can be rewritten to $xF(x) - \int_0^x uf(u)\,du$, and then by partial integration we find

$$\int_0^x f(u)(x-u)\,du = xF(x) - \left(xF(x) - \int_0^x F(u)\,du \right) = \int_0^x F(u)\,du \tag{5.24}$$

Thus we may rewrite the cost equation:

$$C(x) = C_R(x) + (x_O - x)c_r + c_D \int_0^x F(u)\,du + c_F F(x) \tag{5.25}$$

Assuming the gamma process is appropriate for describing the degradation, and given we know the degradation at t_0, i.e., $y_0 = X(t_0)$, the failure threshold ℓ and the degradation parameters α and β we have from Equation (5.20):

$$F_{RUL}(t) = 1 - F_{\alpha t,\beta}(\ell - y_0) \tag{5.26}$$

Where $F_{\alpha t,\beta}(\cdot)$ is the CDF of the gamma distribution with parameters αt and β respectively. Replacing $F()$ in Equation (5.24) by $F_{\text{RUL}}()$ from Equation (5.26) gives:

$$C(x) = C_{\text{R}}(x) + (x_{\text{O}} - x)c_{\text{r}} + c_{\text{D}} \int_0^x \left[1 - F_{\alpha u,\beta}(\ell - y_0)\right] \mathrm{d}u + c_{\text{F}} \left[1 - F_{\alpha x,\beta}(\ell - y_0)\right] \tag{5.27}$$

Problems

5.1 Wind turbine example

We are considering a wind turbine. The wind energy models are presented in Chapter 7, but we introduce the basic idea. The power of a wind turbine is given by [w=Watt]:

$$P = sec : Off W \frac{1}{2} \rho A u_0^3$$

where γ is the yaw angle, $A = \pi(D_{\text{r}}/2)^2$ is the rotor swept area and u_0 is the free-stream wind velocity. The reduction factor is given by:

$$C_{\text{P}}(a, \gamma) = 4a(cos(\gamma) - a)^2$$

where a is the axial induction factor. It is easy to show that the maximum of $C_{\text{P}}(a, \gamma)$ is achieved for $a = a^* = \cos(\gamma)/3$ and the maximum theoretical power is $C_{\text{P}}^* = 16/27 \cos^3(\gamma)$ also known as the Betz limit when $\gamma = 0$.

In this problem we assume the following quantities:

- $u_0 = 10$ # Free wind speed [m/s]
- $\rho = 1.225$ # Air density at 15 degrees [kg/m^3]
- $r = 60$ # Rotor radius [m]
- $D_r = 2 * r$ # Rotor diameter [m]
- $\gamma = 0$ # No yawing
- MTTF $= 8760 * 5$ # Mean time to failure without maintenance [hours] = 5 years
- $\ell = 100$ # Failure limit, normalized to 100 (%)
- $\mu = \ell$/MTTF # Drift parameter for degradation [hours^{-1}]
- $\sigma = 50 * \mu$ # Infinitesimal standard deviation, volatility of degradation [hours^{-1}]

- $T_{\mathrm{L}} = 2 * 7 * 24$ # Lead time of maintenance = 2 weeks
- $c_{\mathrm{R}} = 2\ 500\ 000$ # Cost of renewal [NOKs]
- $c_{\mathrm{F}} = 7\ 500\ 000$ # Cost of failure [NOKs]
- $p_{\mathrm{e}} = 0.5$ # Energy price [NOK/kWh]

a) Calculate the power, and profit per hour = c_U = Loss per hour if not producing

b) Assume mu and sigma are the underlying parameters in the Wiener process. Use μ, σ and ℓ to calculate the mean time to failure and standard deviation of the time to failure

c) Find the optimal maintenance level m, i.e., at which degradation level should renewal be ordered. What is the average total cost per hour (maintenance, failure and production loss)?

d) By allowing longer lead times, i.e., 3 weeks, the cost of renewal could be reduced to $c_R = 2\ 000\ 000$. Would this give a total cost reduction compared to the original problem?

Chapter 6

Markov State Model - An introduction

6.1 Introduction

Consider a stochastic process $\{Y(t), t \in \Theta\}$, where $Y(t)$ describes the state (deterioration level) of an item at time t. In the following we assume that the state variable only takes a finite number of states. We first present the model when no maintenance is carried out, i.e., we start at time $t = 0$ and observe the system until failure. Let:

$$\begin{aligned} Y(0) &= y_0 \\ Y(T) &= y_r \end{aligned} \tag{6.1}$$

where T per definition is the time of the first failure. Between y_0 and y_r there are $r-1$ intermediate sates. By choosing a large value of r we could obtain a very good approximation to a continuous process if this is required. Let $\tilde{T}_i, i = 0, \ldots, r-1$ be sojourn times, i.e., how long the system stay in state i. Notationally we will typically denote the states by their number rather than by the value to simplify notation.

For the initial model we assume that the sojourn times are independent and exponentially distributed with parameter λ_i. Later on we will investigate how sojourn times may be modelled by arbitrary distributions. We also assume that the process runs through all states chronologically from y_0 to y_r without "stepping back" at any time.

Before we present the modelling framework for this simple situation we introduce the maintenance model. Figure 6.1 depicts the development of $Y(t)$ as a function of time. On the x-axis it is indicated that the system is inspected at period of times $\tau, 2\tau, 3\tau, \ldots$. If the system is found in state $Y(t) \geq y_l$ at an inspection, the system is renewed to an as good as new state, i.e., y_0.

We now go back to the simple situation where maintenance is not considered. Let $P_i(t)$ denote the probability that the system is in state i at time t. By standard Markov considerations

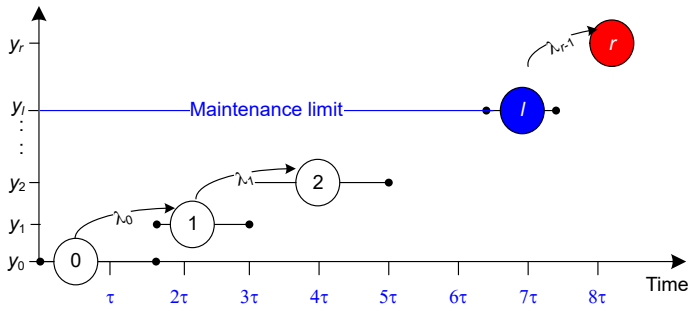


Figure 6.1: Markov transition diagram

we obtain the Markov differential equations:

$$P_i(t+\Delta t) \approx P_i(t)(1-\lambda_i \Delta t) + P_{i-1}(t)\lambda_{i-1}\Delta t \tag{6.2}$$

where Δt is a small time interval and and we set $\lambda_{-1} = 0$ per definition. Further the initial conditions are given by:

$$\begin{aligned} P_0(0) &= 1 \\ P_i(0) &= 0 \text{ for } i > 0 \end{aligned} \tag{6.3}$$

Equation (6.2) could easily be integrated by a computer program, for example VBA in MS Excel. It is now easy to find MTTF by another integration, i.e.,

$$\text{MTTF} = \int_0^\infty R(t)\,dt = \int_0^\infty [1 - P_r(t)]\,dt \tag{6.4}$$

and we should verify that we get MTTF $= \sum_{i=0}^{r-1} \lambda_i^{-1}$. Note that the transition rates, λ_i's, are assumed to be known, that is either they are estimated from data, or found by expert judgement exercises.

6.2 Maintenance model

This section derives a basic maintenance model based on the situation depicted in Figure 6.1. The state variable $Y(t)$ evolves as a function of time. On the x-axis it is indicated that the system is inspected at period of times $\tau, 2\tau, 3\tau, \ldots$. If the the system state at an inspection is equal to, or above the maintenance limit ℓ, then a repair is carried out bringing the system back to state 0. We assume that repair time could be neglected.

The objective function, or cost function to minimize is given by:

$$C(\tau,\ell) = c_{\text{I}}/\tau + (c_{\text{F}} + c_{\text{CM}})\lambda_{\text{E}}(\tau,\ell) + c_{\text{RC}}\rho_{\text{E}}(\tau,\ell) \tag{6.5}$$

where

- ℓ = the maintenance limit
- c_{I} = the cost of an inspection
- c_{F} = the total expected cost of a failure, i.e., downtime cost and trip cost, and any safety cost
- c_{CM} = the cost of repairing a failed item
- c_{RC} = the cost of renewing a degraded item, i.e., not failed but above or equal the maintenance limit
- $\lambda_{\mathrm{E}}(\tau,\ell)$ = the effective failure rate for an item inspected at regular intervals of length τ and renewed if the maintenance limit is reached at an inspection
- $\rho_{\mathrm{E}}(\tau,\ell)$ = expected number of renewals per unit time for an item inspected at regular intervals of length τ and renewed if the maintenance limit is reached at an inspection

In order to specify the model depicted in Figure 6.1 we need to specify the r transition rates $\lambda_0,\lambda_1,\ldots,\lambda_{r-1}$. To simplify the specification we make the following assumption:

- $\lambda_{i+1} = (1+\nu)\lambda_i$, i.e., the transition rates are increasing as the item is degrading, with a constant factor $1+\nu$ corresponding to exponential growth
- It is possible to specify $V = \lambda_{r-1}/\lambda_0$, i.e., how much faster the growth is at end of life compared to initially
- It is possible to specify MTTF, i.e., the mean time to failure if no maintenance is carried out
- There is a fixed probability, q that an inspection will not reveal that the maintenance limit is reached

It is rather easy to show the following:

- $\nu = V^{1/(r-1)}$
- $\lambda_0 = \frac{1-1/\nu^r}{(1-1/\nu)\mathrm{MTTF}}$
- $\mathrm{Var}(T) = \frac{1-\nu^{-2r}}{(1-\nu^{-2})\lambda_0^2}$ = variance of time to failure if no maintenance is carried out

We don't need $\mathrm{Var}(T)$ to proceed, but the expression for the variance could be used to compare our Markov model with e.g., a Weibull model where the variance is already assessed.

6.3 A more general transition model

Equation (6.2) may be used in situations where we only allow transitions from state i to state $i+1$. In more general situations there could be transitions in principle from any state i to any state j. For example there could be two degradation mechanisms, smooth degradation and shocks. The smooth degradation causes jumps from one state to the next, whereas shocks could cause larger jumps. In this situation we need to work with matrices. Let $\mathbf{A}$ be an $(r+1)\times(r+1)$ matrix where element (i,j) represents the constant transition rate from state i to state j. The indexing here starts at 0, e.g., $\mathbf{A}(0,1) = a_{0,1}$ is the transition from state 0 to state 1.

Further, let $\mathbf{P}(t)$ be the time dependent probability vector for the various states defined in $\mathbf{A}$. We now let $\mathbf{P}(t=0) = [1,0,0,\ldots,0]$ to reflect that the system starts in state 0. From standard Markov theory we now need the Markov differential equations, i.e., $\mathbf{P}(t)\cdot\mathbf{A} = \dot{\mathbf{P}}(t)$, from which it follows:

$$\mathbf{P}(t+\Delta t) \approx \mathbf{P}(t)[\mathbf{A}\Delta t + \mathbf{I}] \tag{6.6}$$

where Δt is a small time interval. Equation (D.7) is now used repeatedly to find the time dependent solution for the entire system. This corresponds to integrating Equation (6.2).

We now outline the main principle for working with matrices to find the time dependent solution and other relevant quantities. Assume we have access to a small library of matrix routines:

```
Function mMult(A,B) -> Returns a matrix equal to  A * B
Function fixA(A) -> Fill diagonal of A such that sumrow=0
Function getIntMatrix(A, DeltaT) -> [A * DeltaT + I]
```

In the following we assume that the matrix library is defined by standard indexing, i.e., the first row is denoted row number 1 and so on. A warm up exercise to find MTTF is now:

```
Function getMTTF(A)
fixA A
MTTF = initial guess
DeltaT = MTTF / 1000
hlp = 0
t=0
P=[1,0,0,....]
IM = getIntMatrix(A, DeltaT)
Do While t < 5*MTTF
   P = mMult(P, IM)
```

```
    hlp = hlp +  (1-P(r+1)) * DeltaT
    t = t + DeltaT
Loop
getMTTF = hlp
End Function
```

To get higher precision we could increase the integration to e.g., 10MTTF. Note the motivation for this approach is given by:

$$\text{MTTF} = \int_0^\infty R(t)\,dt = \int_0^\infty [1 - P_r(t)]\,dt \tag{6.7}$$

where $1 - P_r(t)$ is the probability that we are not in state r at time t.

So far the maintenance regime is not reflected in the approach. Let $\lambda_{\text{E}}[(]\tau, l)$ be the effective failure rate, i.e., the expected number of failures per unit time if the system is inspected every τ time unit, and renewed whenever $Y(t) \geq y_l$ at an inspection. In the integration of Equation (D.7) we start with $t = 0$ and whenever t coincides with τ, 2τ etc., special actions are taken:

```
Function lambdaEffective(A,tau,l)
fixA A
MTTF = getMTTF(A)
DeltaT = MTTF / 1000
hlpF = 0
t=0
localTime=0
P=[1,0,0,....]
IM = getIntMatrix(A, DeltaT)
Do While t < 10*MTTF
    P = mMult(P, IM)
    hlpF = hlpF + P(r + 1)        Add to effective failure rate
    P(1) = P(1) + P(r + 1)        If system is failed, it is assumed to be
        renewed
    P(r + 1) = 0                  Clear probability
    If localTime >= tau Then
      sumP = 0
      For i = l+1 To r
          SumP = SumP + P(i)
          P(i)=0
       Next i
```

```
        P(1) = P(1) + SumP
        localTime = 0
    Else
        localTime = localTime + DeltaT
    End If
    t = t + DeltaT
Loop
lambdaEffective = hlpF / t
End Function
```

Note the indexing, i.e., the failed state is $r+1$ and the maintenance limit is $l+1$. In Python indexing starts at =0, so a Python script is even more intuitive.

In the `If localTime = tau` part of the script above we have used a loop to simulate what is happening during an inspection. A more efficient way to do this would be to create an "inspection matrix", say **M** defined by:

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{6.8}$$

where the starting point is an identity matrix, but where we from the row corresponding to state l shift the "ones" to the left.

```
    :
    If localTime >= tau Then
        P = mMult(P, M)
        localTime = 0
    Else
    :
```

Such an inspection matrix could also be used to specify that an inspection is not perfect. For example if q is the probability that an inspection fails to reveal that the actual state is l or higher, the corresponding leftmost "one" is replaced by $1-q$ and the diagonal element is replaced by

q for rows corresponding to states $l, l+1, \ldots, r-1$. An inspection matrix could also be used to specify that upon an inspection it might be decided to repair to a state which is not as good as new. For example in 80% of the cases we repair to state 0, in 15% of the cases we repair to sate 1 and in 5% of the cases we repair to state 2.

6.3.1 Significant repair times

So far we have assumed that repair times could be neglected. If we can not neglect repair times we need to model repair times in the transition matrix **A**. For example if at an inspection we with some probability q will decide to repair from state i to state j with constant repair rate μ a first approach would be to modify the **A**-matrix, i.e., $\mathbf{A}(i,j) = a_{i,j} = q\mu$. However, this would imply that a repair starts immediately after the system has reached state j. In reality, a repair can first start after the coming inspection.

To handle the situation we now introduce "virtual" states. A virtual state is a state in the **A**-matrix representing the situation where a maintenance action has been decided and the repair is actually started. For each pair (i,j) where there could be a repair from physical state i to physical state j a virtual state $k_{i,j}$ is defined. Then the associated transition rate is set to $a_{k_{i,j},j} = \mu$. The row and column representing the virtual state $k_{i,j}$ can be any ones larger than those already "occupied". The inspection matrix **M** will also get an additional row and column representing the virtual state $k_{i,j}$, where $\mathbf{M}(i,k_{i,j}) = q$, where we in addition need to ensure that the row sum equals one.

Note that while repairing from state i to state j represented by $a_{k_{i,j},j} = \mu$ there might be a "competing" transition from for example state i to state l, thus we also need to specify $a_{k_{i,j},l} = \lambda_{i,l}$. Such transitions are not shown in Figure 6.2.

Figure 6.2 illustrates the Markov diagram for a situation with $r = 4$. Here λ_{i_j} is the transition rate from state i to state j representing degradation. Further $\mu_{k_{i,j},j}$ is the repair rate from virtual state $k_{i,j}$ to state j. When a repair is initiated as a result of a proof-test, virtual states are introduced. For example the state (2.1) represent that after a test it is decided to repair from state 2 to state 1. The doted lines represent transitions that instantaneously take place after a proof-test. The probabilities given by the q-values represent maintenance decisions. For example $q_{3,3.0} = 1$ represents that if a state 3 is revealed by a proof-test, we always initiate a repair to state 0. $q_{2,2.0}$ is representing the probability that we after revealing a state 2 on a proof-test we initiate a repair to state 0. The q-values are entered into the inspection matrix, **M**.

In Figure 6.2 there are three nodes representing that the system is in a "small degradation" state, i.e., physical state 2. State 2 in the diagram is a *hidden state*, we are not aware of any transition from state 1 to state 2. The states (2.0) and (2.1) are *evident states*, we know that we are in main state 2 (small degradation), a maintenance request has been issues (to state 0 and state 1 respectively).

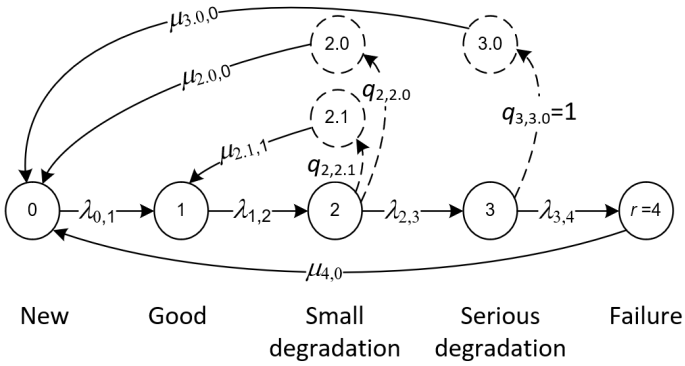


Figure 6.2: Markov transition diagram with potential repairs

Note that in Figure 6.2 we use the notation $a_{\text{From,To}}$ without indicating the actual row and column *numbers* in the transition matrix. The notation $a_{k_{i,j},l}$ on the other hand, is used to identify a row and column number in a matrix in the code when we do the programming.

In previous sections we have focused on the effective failure rate, but we might also be interested in the average portion of time we are in each state. For example we may use:

```
:
Do While t < 10*MTTF
    P = mMult(P, IM)
    Pavg = Pavg + P
    If localTime >= tau Then
        P = mMult(P, M)
        localTime = 0
    Else
        localTime = localTime + DeltaT
    End If
    t = t + DeltaT
Loop
Pavg = Pavg * DeltaT / t
:
```

Phase type modelling

So far we have assumed that the sojourn times are exponentially distributed. This assumption could be questioned if there are failure mechanisms like wear, fatigue, corrosion etc. that drives the degradation of the system. Phase type modelling is an approach where we may approximate a stochastic variable with a multi state Markov model. The more states we use the better will the

approximation be. In the following we do not discuss the explicit fitting of model parameters for the approximation. Several statistical packages exist for this. We will also assume that each random variable is approximated by a two-state Markov model in order to reduce the total number of states. See Laskowska and Vatn (2020) for an example using phase type modelling.

Model

Consider a system having three main states, 1 = new, 2 = degraded, and F = failed. With *main state* we here mean what the categorization used by the maintenance department.

We assume that sojourn times in state 1 and state 2 could be approximated by a phase type distribution, i.e., $\tilde{T}_1 \sim F_1(t)$ and $\tilde{T}_2 \sim F_2(t)$. The sojourn times are assumed to be stochastically independent.

For $\tilde{T}_1$ the phase type model comprises two sub states, i.e., $1a$ and $1b$. A acyclic phase type model is used where:

- The probability that the system starts in sub state $1a$ is p_1, and the probability that the system starts in sub state $1b$ is $1-p_1$
- There is a constant transition rate, λ_{1a} from state $1a$ to $1b$
- There is a constant transition rate, λ_{1b} from state $1b$ to an absorbing state outside the system

A similar model exist for sojourn time 2.

Note that given $F_1(t)$ and $F_2(t)$ we may in principle find the distribution of the time to failure for this system by the convolution theorem. This will not be pursued here.

System modelling

The phase type models for the two sojourn time enable easy integration by use of the Markov equations $\mathbf{P}(t+\Delta t) \approx \mathbf{P}(t)[\mathbf{A}\Delta t + \mathbf{I}]$ when each sojourn time is treated independently.

In order to have a complete model we need to link the two phase type models.

Proposition: For the phase type distribution approximating the first sojourn time there is a rate λ_{1b} from state $1b$ to an absorbing state outside the system. This transition is split into two transitions, one to state $2a$ and one to state $2b$. The corresponding rates are $\lambda_{1b\_2a} = p_2\lambda_{1b}$ and $\lambda_{1b\_2b} = (1-p_2)\lambda_{1b}$ respectively. Figure 6.3 depicts the situation:

The dashed ellipses represent the physical or main states “new” and “degraded” as observed by, e.g., maintenance personnel. The states $1a$, $1b$, $2a$ and $2b$ are artificial or sub states used for modelling, but have no physical interpretation.

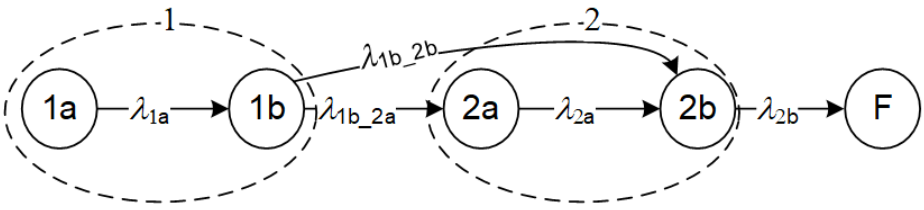


Figure 6.3: Markov transition diagram with intermediate transitions

Let $P_F(t)$ be the probability that the system is in state F at time t. $P_F(t)$ can easily be obtained by integration the Markov equation for the compound system yielding the cumulative distribution function for time to system failure.

For the model with the artificial states, the Markov property holds. However, if we consider the three main states (1=new, 2=degraded, and F=failed), the Markov property does not hold. Given that we have stayed in main state 2 for some time, the probability of escaping that state will typically increase as time goes by. The explanation is that given we entered state 2 at time x, we either went to the artificial state $2a$ with probability p or to state $2b$ with probability $1-p$, which is exactly matching the phase type model we use for the "physical" state 2 when the "failure rate function" for state 2 is increasing.

6.4 Varying intervals for inspection

The idea of having fixed lengths between inspections may be questioned. Obvious it would be some administrative advantages if we could stick to the same inspection intervals independent on the system state. On the other hand it seems reasonable to reduce the inspection intervals as we approach the maintenance limit. To obtain the total failure rate, number of inspections, and number of repairs for a general inspection and maintenance strategy is almost impossible. In the following we present an approach where we make some assumptions which would be reasonable to handle from an administrative point of view, and which also should not be far from the optimal solution:

- The time intervals between inspections are either τ_L (long), τ_M (medium), or τ_S (short). Further k_L and k_M are integers such that $\tau_L = k_L\tau_S$ and $\tau_M = k_M\tau_S$.
- $\mathscr{L}_L$ is a set of states which require inspection interval τ_L, $\mathscr{L}_M$ is a set of states which require inspection interval τ_M and $\mathscr{L}_S$ is a set of states which require inspection interval τ_S. For all other states, $\mathscr{L}_R$, it is required to repair the system to a good as new state.
- If a failure occurs at time t in between inspections the system will be repaired to a good as new condition immediately, and the first inspection of length τ_L will take place at the

smallest value of $k\tau_S$ where k is an integer and $k\tau_S > t + \tau_L$.

The mathematical formulation of this problem is rather difficult, and is beyond the scope for the course HAV6003. However, the main approach is presented for the interested reader.

In the modelling we assume that we at any time, $t \neq k\tau_S$, have:

- The current inspection regime is governed by τ_L, τ_M or τ_S
- Let t_C be the starting point of the current inspection interval, i.e., $t_C > k\tau_S$ and $t_C < (k+1)\tau_S$ for some integer k
- If the current inspection regime is governed by τ_L, the next inspection will take place at one of the following point of times: $t_C + \tau_S, t_C + 2\tau_S, \ldots, t_C + k_L\tau_S$
- If the current inspection regime is governed by τ_M, the next inspection will take place at one of the following point of times: $t_C + \tau_S, t_C + 2\tau_S, \ldots, t_C + k_M\tau_S$

We may now define different $\mathbf{P}$-vectors to hold the time dependent state probabilities as we integrate the solution. Let $\mathbf{P}_S(t)$ be defined such that

$$P_{i,S}(t) = \Pr(Y(t) = y_i \cap \text{ current regime is } \tau_S) \tag{6.9}$$

For the medium and long intervals we also need to take into account the "starting point" of these, and we define:

$$P_{i,L,m}(t) = \Pr(Y(t) = y_i \cap \text{ current regime is } \tau_L \cap \text{ cycle is } m) \tag{6.10}$$

With "cycle" m we mean that the inspection will take place at point of times $\tau_L + (m-1)\tau_S, 2\tau_L + (m-1)\tau_S, 3\tau_L + (m-1)\tau_S, \ldots$. We can imagine that these cycles are running in parallel. In reality it will be only one cycle that could be active, but which one is actually active depends on when the system is renewed. Similarly we define:

$$P_{i,M,n}(t) = \Pr(Y(t) = y_i \cap \text{ current regime is } \tau_M \cap \text{ cycle is } n) \tag{6.11}$$

In total we have one $\mathbf{P}_S(t)$-vector, k_M $\mathbf{P}_{M,n}(t)$-vectors and k_L $\mathbf{P}_{L,m}(t)$-vectors. As we integrate the Markov differential equations for $t \neq l\tau_S$ we update all the $\mathbf{P}(t)$-vectors according to Equation (D.7).

Here it should be noted that a more efficient approach would be to use:

$$\mathbf{P}(t + \tau_S) \approx \mathbf{P}(t)[\mathbf{A}\Delta t + \mathbf{I}]^{2^n} \tag{6.12}$$

where $\Delta t = \tau_S/2^n$ and n is sufficient large to get a low value of Δt. Typically $n = 10$. This means that we may calculate $[\mathbf{A}\Delta t + \mathbf{I}]^{2^n}$ once, and use this matrix for all integrations. An alternative is to use matrix exponentials if we have the numerical routine available, i.e., $\mathbf{P}(t+\tau_S) = \mathbf{P}(t)e^{\mathbf{A}\tau_S}$.

Initially we have $P_{0,L,1}(t=0)$, and all other probabilities are equal to zero. We now apply Equation (6.12) for $t = 0, \tau_S, 2\tau_S, 3\tau_S, \ldots$ for all the $\mathbf{P}(t)$-vectors. At each step we investigate each $\mathbf{P}(t)$-vector with respect to:

- Count the "number" of failures, to update the effective failure rate λ_E
- Count the "number" of repairs, i.e., update the renewal rate ρ_E
- Move "probability mass" to reflect repairs and change of inspection regime

For each $\mathbf{P}(t)$-vector we investigate the element corresponding to the failed state. These probabilities are added to a variable holding the accumulated expected number of failures. A failure could have occurred anywhere in the interval we integrate by Equation (6.12), and we assume an immediate repair. However, according to our assumptions, the system will not change the point of times where inspections are possible. Now consider time $t = l\tau_S$, then there will be a maintenance regime, say m with inspection interval τ_L which also has an inspection at time $t = l\tau_S$. Let p_j be all the probabilities representing failures in the set of $\mathbf{P}(t)$-vectors. The probabilities are now moved according to:

$$P_{0,L,m}(t^+) = P_{0,L,m}(t^-) + \sum_j p_j \tag{6.13}$$

where we assume that τ_L is small. Since the failure could have taken place some time before t, it is a probability that the system was reset to an as good as new state, and then jumped to the next deterioration level if τ_L is large. If this is the case we could split $\sum_j p_j$ to state 0 and 1.

We now proceed to handle the *change* of maintenance regime. As before we identify the integer value m which is such that the regime τ_L with cycle m has due date for an inspection at time $t = l\tau_S$. Similarly we identify the integer value n which is such that the regime τ_M with cycle n has due date for an inspection at time $t = l\tau_S$.

To understand the situation, consider $\tau_L = 6, \tau_M = 3$ and $\tau_L = 1$. Assume we are considering an inspection at time $t = l\tau_S = 13 \cdot 1 = 13$. m will now be 2 since the second τ_L cycle will have an inspection at times $1, 7, \mathbf{13}, 19\ldots$. Further $n = 2$ because the second τ_M cycle will have an inspection at times $1, 4, 7, 10, \mathbf{13}, 16\ldots$.

First consider $\mathbf{P}_{L,m}(t)$, i.e., the vector representing cycle m for the regime τ_L. For all states $i \in \mathscr{L}_L$ there will be no change in the inspection regime. For all states $i \in \mathscr{L}_M$ this will correspond to shifting from regime τ_L to regime τ_M. The vector $\mathbf{P}_{M,n}(t)$ is now representing the cycle which will "take over". Further, for all states $i \in \mathscr{L}_S$ this will correspond to shifting from regime τ_L to

regime τ_S. Finally for all states $i \in \mathscr{L}_R$ this will correspond to a repair. We thus have:

$$\begin{aligned} P_{i,M,n}(t^+) &= P_{i,M,n}(t^-) + P_{i,L,m}(t^-), i \in \mathscr{L}_M \\ P_{i,S}(t^+) &= P_{i,S}(t^-) + P_{i,L,m}(t^-), i \in \mathscr{L}_S \\ P_{0,L,m}(t^+) &= P_{0,L,m}(t^-) + \sum_{i \in \mathscr{L}_R} P_{i,L,m}(t^-) \\ P_{i,L,m}(t^+) &= 0, i \notin \mathscr{L}_L \end{aligned} \tag{6.14}$$

The notation t^- and t^+ is used to denote time just before and just after an inspection respectively.

Referring to the example this means that if there is an inspection at time $t = 13$ there is a τ_L regime with cycle $m = 2$ which has an inspection at time $t = 13$. If it during the inspection is observed that the system is in $\mathscr{L}_M$, i.e., we find positive probabilities for $P_{i,L,m=2}(t^-), i \in \mathscr{L}_M$ we shift to a τ_L regime with cycle $n = 2$, i.e., inspection at time $t = 13$ and next inspection time at $t = 13 + 3 = 16$.

Next consider $\mathbf{P}_{M,n}(t)$. For all states $i \in \mathscr{L}_M$ there will be no change in the inspection regime. For all states $i \in \mathscr{L}_S$ this will correspond to shifting from regime τ_M to regime τ_S. The vector $\mathbf{P}_S(t)$ is now representing the regime which will "take over", i.e.,

$$\begin{aligned} P_{i,S}(t^+) &= P_{i,S}(t^-) + P_{i,M,n}(t^-), i \in \mathscr{L}_S \\ P_{0,L,m}(t^+) &= P_{0,L,m}(t^-) + \sum_{i \in \mathscr{L}_R} P_{i,M,n}(t^-) \\ P_{i,M,m}(t^+) &= 0, i \notin \mathscr{L}_M \end{aligned} \tag{6.15}$$

Finally consider $\mathbf{P}_S(t)$. For all states $i \in \mathscr{L}_S$ there will be no change in the inspection regime. For all states other states this will correspond to a repair to an good as new state. Note that for this regime there is not possible to be in $\mathscr{L}_L$ or $\mathscr{L}_M$. The updating of probabilities is defined by:

$$\begin{aligned} P_{0,L,m}(t^+) &= P_{0,L,m}(t^-) + \sum_{i \in \mathscr{L}_R} P_{i,S}(t^-) \\ P_{i,S}(t^+) &= 0, i \in \mathscr{L}_R \end{aligned} \tag{6.16}$$

6.5 Varying intervals for inspection - Alternative approach

In the previous section the Markov differential equations were integrated having all possible combination of sequences in the various $\mathbf{P}(t)$ vectors. Alternatively we could integrate the differential equations but when there is a failure or a demand for a renewal we just remove the corresponding probability mass from the $\mathbf{P}(t)$ vectors. This will reduce the number of $\mathbf{P}(t)$ vectors to consider.

We still assume that the time intervals between inspections are either τ_{L}, τ_{M}, or τ_{S}. Further k_{L} and k_{M} are integers such that $\tau_{\mathrm{L}} = k_{\mathrm{L}}\tau_{\mathrm{S}}$ and $\tau_{\mathrm{M}} = k_{\mathrm{M}}\tau_{\mathrm{S}}$. The situation now simplifies because there is only one τ_{L} regime and one τ_{M} regime, i.e., we are not considering the cycles any more. The time dependent probability vectors for each regime is given by $\mathbf{P}_{\mathrm{L}}(t)$, $\mathbf{P}_{\mathrm{M}}(t)$ and $\mathbf{P}_{\mathrm{S}}(t)$.

In the integration we now define $\mathbf{h}$ to be a vector of probabilities of a failure in each period of length τ_{S}. Similarly $\mathbf{g}$ is a vector of probabilities of a request of a renewal in each period. Let j be an index running through all intervals of length τ_{S}. The $\mathbf{P}$-vector elements are defined by $P_{0,\mathrm{L}}(t=0) = 1$ and 0 for all other elements in the set of $\mathbf{P}(t)$-vectors. The integration procedure is now:
For $j = 1, 2, \ldots$ integrate all $\mathbf{P}(t)$-vectors according to:

$$\mathbf{P}(t+\tau_{\mathrm{S}}) \approx \mathbf{P}(t)[\mathbf{A}\Delta t + \mathbf{I}]^{2^n} \tag{6.17}$$

where $\Delta t = \tau_{\mathrm{S}}/2^n$, and where we update $t = t + \tau_{\mathrm{S}}$. In the formulas that follow t^- represents the time just prior to t, and t^+ represents the time just after t, i.e., when adjusting for the decision to take at time t.

Collect failure probabilities etc.:

$$h(j) = \sum_{k\in\{\mathrm{L,M,S}\}} P_{r,k}(t^-) \tag{6.18}$$

$$g(j) = \sum_{i\in\mathscr{L}_{\mathrm{R}}} P_{i,\mathrm{S}}(t^-), \text{ if } j \geq k_L \tag{6.19}$$

If $(j-1) \mod k_M = 0$ and $j > k_L$ then ("medium" maintenance):

$$g(j) = g(j) + \sum_{i\in\mathscr{L}_{\mathrm{R}}} P_{i,\mathrm{M}}(t^-) \tag{6.20}$$

$$P_{i,\mathrm{S}}(t^+) = P_{i,\mathrm{S}}(t^-) + P_{i,\mathrm{M}}(t^-), i \in \mathscr{L}_{\mathrm{S}} \tag{6.21}$$

End If
If $(j-1) \mod k_L = 0$ then ("long term" maintenance):

$$g(j) = g(j) + \sum_{i\in\mathscr{L}_{\mathrm{R}}} P_{i,\mathrm{L}}(t^-) \tag{6.22}$$

$$\begin{aligned} P_{i,\mathrm{S}}(t^+) &= P_{i,\mathrm{S}}(t^-) + P_{i,\mathrm{L}}(t^-), i \in \mathscr{L}_{\mathrm{S}} \\ P_{i,\mathrm{M}}(t^+) &= P_{i,\mathrm{M}}(t^-) + P_{i,\mathrm{L}}(t^-), i \in \mathscr{L}_{\mathrm{M}} \end{aligned} \tag{6.23}$$

End If
In the procedure listed above we have not explicitly "removed" the probability mass corresponding to a "request" for renewal. This we have to do.

The vectors **h** and **g** represent the expected number of failures and the number of requested renewals for each interval. Let **w** be the vector of the total expected number of renewals in each interval. For interval number 1 we have $w(1) = g(1)$, for interval number 2 the expected number of renewals is $w(2) = h(2) + w(1)g(1)$. To find the number of expected renewals in general we could use the discrete version of the renewal density for period j:

$$w(j) = g(j) + \sum_{i=1}^{j-1} w(j-i)g(i) \tag{6.24}$$

Since we already have calculated the values in **g** it is straight forward to obtain **w** from Equation (6.24).

Note that $w(j-i)g(i)$ represents the probability that there was a renewal at $(j-i)\tau_S$ and then there is another renewal $i\tau_S$ later. Here we should also account for the possibility that it was a *failure* at $(j-i)\tau_S$ and then there is another renewal $i\tau_S$ time units later, hence more correct would be:

$$w(j) = g(j) + \sum_{i=1}^{j-1} \left[w(j-i) + f(j-i) \right] g(i) \tag{6.25}$$

where $f()$ is described below.

Let **f** be the vector of expected number of failures in each period. A failure will occur in period j in two disjoint ways, either the initial system fails in interval j, or there was a renewal or failure in a previous period $j-i$, and then this system fails after another i periods:

$$f(j) = h(j) + \sum_{i=1}^{j-1} \left[w(j-i) + f(j-i) \right] h(i) \tag{6.26}$$

Equation (6.26) may now be used to find the average effective failure rate over a given time horizon.

6.6 Basic Markov degradation model

The ISO standard for collection and exchange of reliability and maintenance data for equipment ISO14224 (2016) proposes three different levels of degradation of a components:

- Incipient (I): The item is able to perform it's required function. Some degradation is observed, but it is not expected to result in a failure in short time.

- Degraded (D): The item is able to perform it's required function, but with possible reduced performance, and a failure is expected in rather short time if a corrective action is not carried out.
- Critical (C): The item is not able to perform it's required function, i.e., the item is in a fault state.

Many operating companies in the oil- and gas industry has adopted these levels in both their reporting systems, i.e., computerized maintenance management systems (CMMS) as well as when it comes to planning and optimization of the maintenance strategies. In the following we will there fore adapt these categories. In order to have a complete set of states for our Markov model we make two adjustments:

1. A main category *healthy* (H) is introduced to represent a component which is considered to be as good as new, or almost as good as new
2. For each of the main categories (main states) H, I and D, we introduce two sub-states to allow a for a refinement for the description. These extra sub-states are labelled + and - to represent a state slightly better and a state slightly worse than average of the main state.

The use of sub-states serves two purposes. First of all it will allow a more realistic degradation modelling. Since usually we assume physical degradation of the item under consideration, a Markov model with only two states between a perfect component and a failed component is not very realistic. Secondly, it will allow maintenance personnel to bring more evidence of the situation into the mark, and finally the maintenance department could establish more flexible maintenance strategies. Typically observing a degraded state, i.e., D, according to the categories will result in a decision to do a corrective action to bring the item into a good as new state. With the refined codes, it is possible to trigger a maintenance action in the D+ state for very critical items, whereas for non-critical items such a maintenance limit could be set to D-.

The states are shown in Figure 6.4. We will later discuss how to estimate the transition rates between the states.

To model repair we make the following assumptions:

- We assume that the system is inspected every τ time unit.
- If the result of the inspection is that the condition (state of the system) is equal or higher than the maintenance limit, a maintenance request is made.
- The lead time, i.e.,the time from a maintenance order is placed until it is executed is a random quantity with mean value equal to MLD (Mean lead time delay/Mean logistic delay)

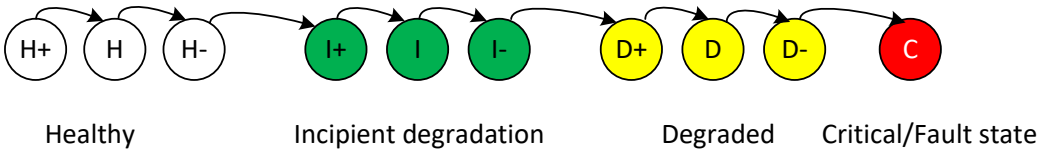


Figure 6.4: Markov transition diagram for the basic situation

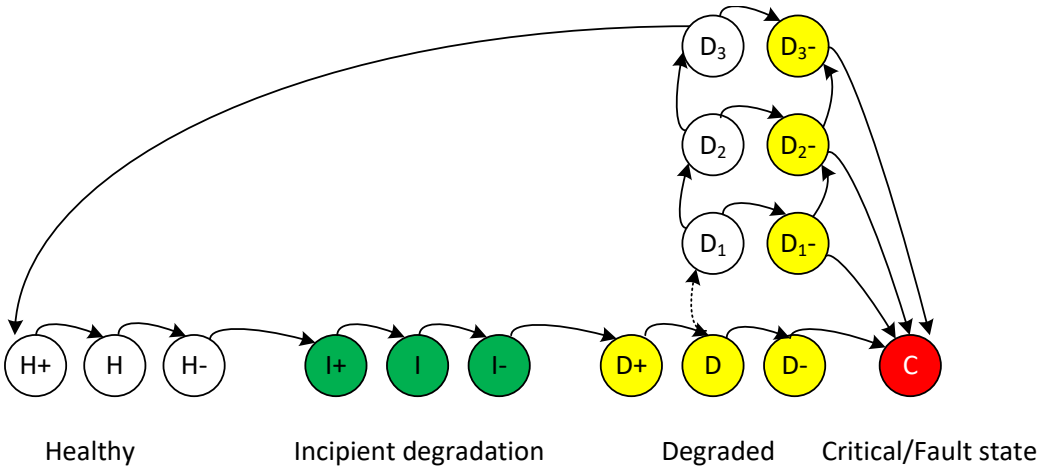


Figure 6.5: Markov transition diagram with lead times

- For simplicity we assume that the active repair time is small compared to MLD, and therefore can be ignored.
- To simplify modelling we assume that lead times are Erlang distributed with shape parameter k. Figure 6.5 depicts the situation when $k = 4$, where dashed states represent that a maintenance request has been placed, but is still pending.
- The degradation process continues after a maintenance request has been placed. The diagram illustrates this by transition rightwards in the dashed states in the diagram.

Note that the transition between state D and D1 is shown by a dotted line. This is because such a transition only takes place at each inspection. The same apply for the transition between state D- and D1-.

To solve the model the following steps are required

1. The system starts in state H+
2. The system is integrated for a long time, i.e., such that we can achieve asymptotic results. Let $\mathbf{P}(t)$ be the time dependent probability vector at time t, and then we apply Let $\mathbf{P}(t + dt) = \mathbf{P}(t)\,[I + \mathbf{A}dt]$

Problems

6.1 Integration of Markov equations. Assume $r = 5$ and $\lambda_i = 0.01, i = 0, 1, \ldots$. Integrate the Markov differential equations and obtain the expected value and variance of the time to failure. Hint: Use partial integration for the variance similar to MTTF $= \int R(t)dt$.

6.2 Effective failure rate and renewal rate. Write a program to calculate the effective failure rate and the renewal rate.

6.3 Wind turbine example with observable degradation. Consider the example in section 4.1.15. Assume that it is possible to observe the degradation by inspections. Assume that a Markov state model with $r = 8$ is a reasonable model in this situation. The inspection cost is assumed to be $c_I = 2\,000$ NOKs, and the failure probability of an inspection is $q = 0.1$. The degradation speed at end of life is considered to be 3 times faster than initially. Find the optimal value of τ and $\ell =$ maintenance limit.

6.4 Write a Python program to implement the mean time to failure (MTTF) indicated in Section 6.3.

6.5 Assume $r = 5$ and $\lambda_i = 0.01, i = 0, 1, \ldots$ (time unit weeks). Assume the system is inspected with intervals of length $\tau = 26$. If the system is found in state $Y(k\tau) = 4$ the system will be renewed. Renewal takes place immediately. The probability that an inspection reveals that the system is in state 4 is 70% when this is the case. Find the effective failure rate for this situation.

6.6 Assume $r = 5$ and $\lambda_i = 0.01, i = 0, 1, \ldots$ (time unit weeks). Assume the system is inspected with intervals of length $\tau = 26$. If the system is found in state $Y(k\tau) = 4$ the system will be renewed. There is a *logistic delay* of in average 4 weeks before the repair takes place. Delay time is assumed to be exponentially distributed. The probability of revealing state 4 is still 70%. Find the effective failure rate for this situation.

6.7 Numerical precision. Assume $r = 5$ and $\lambda_i = 0.01, i = 0, 1, \ldots$ (time unit weeks). Assume the system is inspected with intervals of length $\tau = 26$, i.e., not varying intervals. Assume that the system is in state 0 at time $t = 0$. Find $\mathbf{P}(\tau)$ by using Equation (6.12) when using $n = 4, 6, 8$ and 10. What would be a reasonable value of n.

6.8 Implement the model above, where $r = 5$, $\lambda_i = 0.01$, the maintenance rule is $\tau_L = 52, \tau_M = 26, \tau_S = 18, \mathscr{L}_L = \{0\}, \mathscr{L}_M = \{1, 2\}, \mathscr{L}_S = \{3\}$ and $\mathscr{L}_R = \{4\}$.

6.9 Equinor model. Equinor uses a reporting system with the following values of the health indicator:

H Healthy

U Unwell

S Sick

D Dead

The transition vector is given by:

$$\boldsymbol{\lambda} = [1/15000, 1/8000, 1/3000]$$

where the transitions are H → U, U → S and S → D respectively. The time unit is hours.

a) Verify by numerical integration of the Markov equations that MTTF = 15000 + 8000 + 3000

b) Find the effective failure rate and the effective renewal rate if the inspection interval is $\tau = 2$ months. The maintenance limit is"Sick", i.e., when an inspection reveals the state S the item is set back to a good as new state. We assume that the lead time is negligible.

6.10 Varying intervals. In this problem we will investigate the model in section 6.4. We use a model with $r+1 = 4+1 = 5$ states, where 0 = as good as new, 1 = small defect, 2 = medium defect, 3 = large defect and $r = 4$ = fault state. The transition vector is given by:

$$\boldsymbol{\lambda} = [\lambda_0, \lambda_1, \lambda_2, \lambda_3] = [1/48, 1/36, 1/24, 1/12]$$

where the time unit is months. The maintenance strategy is determined by:

- $\mathscr{L}_{\mathrm{L}} = \{0\}$ = Set of states requiring long intervals
- $\mathscr{L}_{\mathrm{M}} = \{1\}$ = Set of states requiring medium intervals
- $\mathscr{L}_{\mathrm{S}} = \{2\}$ = Set of states requiring short intervals
- $\mathscr{L}_{\mathrm{R}} = \{3\}$ = Set of states where renewal is required
- $\tau_{\mathrm{S}} = 1$ = Shortest interval
- $k_{\mathrm{M}} = 6$ = Medium interval is every k_{M} time relatively to τ_{S}
- $k_{\mathrm{L}} = 12$ = Long interval is every k_{L} time relatively to τ_{S}

Obtain the effective failure rate λ_{E}, the renewal rate ρ_{E}, and the inspection rate i_{E}.

6.11 We consider the situation in Problem 6.10 but will investigate the gain of having intervals depending on the condition. The following cost elements shall be used:

- $c_{\mathrm{I}} = 1$ = cost of inspection
- $c_{\mathrm{R}} = 10$ = cost of renewal

- $c_F = 50$ = cost of failure

a) Obtain the optimal inspection interval τ if system is renewed when an inspection reveals that the system is in state 3, and the inspection interval is the same for all degradation levels.

b) Find the total expected cost per time unit in a)

c) Find the optimal value of τ_S when $k_M = 6$ and $k_L = 12$

d) Find the expected cost in c) and compare the result with b).

Chapter 7

Offshore Wind Modelling

7.1 Introduction

This chapter presents models applicable for maintenance modelling of offshore wind farms. First we introduce a simple model for the power that can be extracted from the wind and thus contribute to the energy production, then we explain the concept of wakes and the negative effects of wakes. Wind farm control is essentially about understanding how wakes can be controlled to obtain a high total energy production from the wind farm and at the same time reduce the load from wakes on the turbines.

7.2 Energy in the wind

The rotor blades convert the momentum of a wind field into aerodynamic forces that drive the rotor. The torque from the rotor is transferred to the generator shaft through the drivetrain. The generator finally converts rotational kinetic power into electrical power. To control the power production and forces on the wind turbine the following control variables are typically available Boersma et al. (2017), see Figure 7.1:

θ: Blade pitch angle - The rotor blades can rotate, with their axis of rotation aligned with the blades, using hydraulic actuators or servo pitch motors. Pitch control can be used to influence the power capture and the loads to the wind turbine.

τ_g: Generator torque - The generator converts mechanical power into electricity. Torque control is used to control the power capture.

γ: Yaw angle - The nacelle can rotate, with the axis of rotation aligned with the tower, using a yaw motor. The yaw angle is defined as the angle between the axial rotor axis and the incoming wind direction.

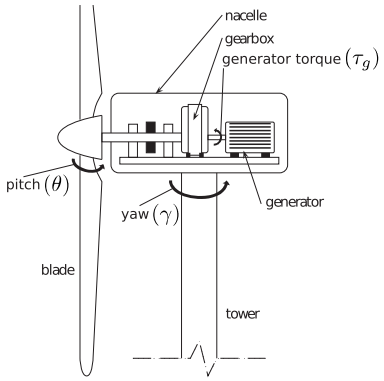


Figure 7.1: Horizontal-axis wind turbine with labelled main components and control variables (Boersma et al., 2017).

Figure 7.2 depicts the free-stream wind with velocity u_0 approaching the wind turbine with rotor diameter D_r yielding a reduced wind speed u_r.

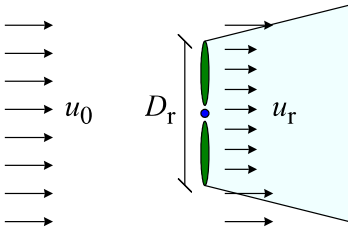


Figure 7.2: Wind energy captured by the turbine

The power that can be extracted from the wind is typically measured in megawatt (MW), and is given by (Boersma et al., 2017):

$$P = C_\mathrm{P}(\theta, \lambda, \gamma)\frac{1}{2}\rho A u_0^3 \tag{7.1}$$

where θ is the blade pitch angle, λ is tip-speed ratio, γ is the yaw angle, $A = \pi(D_\mathrm{r}/2)^2$ is the rotor swept area and u_0 is the free-stream wind velocity.

The reduction factor $C_\mathrm{P}(\theta, \lambda, \gamma)$ depends on quantities we can control by the pitch angle, the yaw angle and the generator torque. To implement a control strategy a mathematical model for $C_\mathrm{P}(\theta, \lambda, \gamma)$ is required, but for our purpose it is sufficient to introduced a simplified reduction factor, i.e.,(Boersma et al., 2017):

$$C_\mathrm{P}(a, \gamma) = 4a(cos(\gamma) - a)^2 \tag{7.2}$$

where a is the axial induction factor given by:

$$a = \frac{u_0 - u_r}{u_0} \tag{7.3}$$

For a given yaw angle γ it is easy to show that the maximum of $C_P(a,\gamma)$ is achieved for $a = a^* = \cos(\gamma)/3$ and the maximum theoretical power is $C_P^* = 16/27\cos^3(\gamma)$ also known as the Betz limit when $\gamma = 0$.

We emphasize again that how to use the control variables (θ, τ_g and γ) is the challenge of wind farm control, but for our purpose it is sufficient to know that this is possible, i.e., to obtain $a^* = \cos(\gamma)/3$ yielding the maximum power $C_P^* = 16/27\cos^3(\gamma)$ for a given γ. Intuitively we should let $\gamma = 0$, meaning that the wind is perpendicular to the turbine swept area, but this will then give wake effects downstream reducing the wind velocity and yielding turbulence to the downstream turbines. Therefore we also need to discuss the concept of wind turbine wakes.

7.3 Wind turbine wakes

Wind turbine wakes are important in offshore wind maintenance because they decrease the power production and increase the loading of downstream wind turbines. Much research has been conducted to understand wind turbine wakes. There exist simple models and more complicated models. In this presentation we only consider a simple model, i.e., the Jensen wake model, Jensen (1983). It will do for our purpose. The motivation for the model is the control volume downstream of the turbine shown in Figure 7.3.

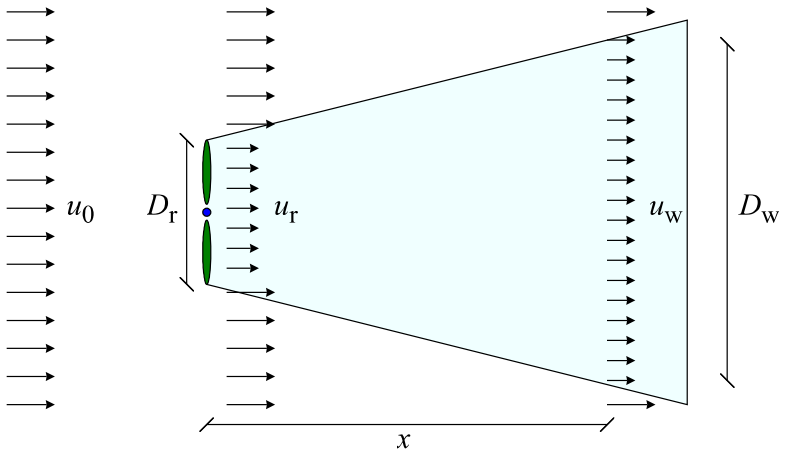


Figure 7.3: The control volume of the Jensen wake model, (Jensen, 1983)

In the model it is assumed that the wake expands linearly downstream the turbine, and at a

distance x downstream the turbine the diameter of the wake is given by

$$D_w = D_r + 2\alpha x \tag{7.4}$$

The decay constant α determines how quickly the wake expands with distance x from the wind turbine. There is some theoretical and experimental research to determine the numerical value of the decay constant which is beyond the scope of this presentation. But it is common to distinguish between onshore and offshore wind, and for offshore it is often recommended to use $\alpha = 0.04$ (Shakoor et al., 2016).

Using a mass balance between the rotor plane and the downstream flow in Figure 7.3 we may write:

$$\rho\pi\left(D_r/2\right)^2 u_r + \rho\pi\left[\left(D_w/2\right)^2 - \left(D_r/2\right)^2\right] u_0 = \rho\pi\left(D_w/2\right)^2 u_w \tag{7.5}$$

assuming air density ρ is constant.

According to Betz theory the value of u_r is given by (Shakoor et al., 2016):

$$u_r = (1-2a)\,u_0 \tag{7.6}$$

where a still is the axial flow induction coefficient.

The general Jensen formula for the wake velocity at distance x for a single wind turbine is found by combining equations (7.4), (7.6) and (7.5):

$$u_w = u_w(x) = u_0\left(1 - 2a\left(\frac{D_r}{D_r + 2\alpha x}\right)^2\right) \tag{7.7}$$

Assuming ideal axially symmetric flow, no rotation, no turbulence and conic shape wake profile, the axial induction factor can also be written as (Göçmen et al., 2016):

$$a = \frac{1 - \sqrt{1 - C_T(a,\gamma)}}{2} \tag{7.8}$$

where the thrust coefficient is given by Katic et al. (1986):

$$C_T(a,\gamma) = 4a(\cos(\gamma) - a) \tag{7.9}$$

and substituted into Equation (7.7) gives:

$$u_w = u_w(x) = u_0\left(1 - \left(1 - \sqrt{1 - C_T(a,\gamma)}\right)\left(\frac{D_r}{D_r + 2\alpha x}\right)^2\right) \tag{7.10}$$

Assuming that the reduction factor $C_P(\theta,\lambda,\gamma)$ is maximized, i.e., the axial induction factor is

$a = a^* = \cos(\gamma)/3$ for a given yawing angle γ, we may use:

$$C_T(a*, \gamma) = C_T(\gamma) = \frac{8}{9}\cos^2(\gamma) \tag{7.11}$$

yielding the following expression for the downstream wind speed at distance x:

$$u_w(\gamma, x) = u_0\left(1 - \left(1 - \sqrt{1 - \frac{8}{9}\cos^2(\gamma)}\right)\left(\frac{D_r}{D_r + 2\alpha x}\right)^2\right) \tag{7.12}$$

Note that in a design phase the objective is to optimize the layout in terms of determine the distance between the turbines, i.e., x, whereas in the operation and maintenance phase x is fixed, and we optimize wrt. the yawing angle γ.

7.4 Direction of wind turbine wakes

The aim of yawing is to direct the wake away from the column of turbines downstream the yawed turbine. Jimenez et al. (2009) have provided an analytical formula for the centre of the downstream wake given by:

$$\phi(x, \gamma) = \frac{dy}{dx} = \frac{\cos^2(\gamma)\sin(\gamma)}{C_T(a, \gamma)}(1 + \beta x)^2 \tag{7.13}$$

where ϕ is the wake skew angle, $\beta = 2\alpha$ is the wake expansion factor. Howland et al. (2016) integrate Equation (7.13) in x to obtain the centre point of the wake in the y-direction at the downstream distance x from the wind turbine and obtain:

$$y_C(x, \gamma) = D_r\frac{\cos^2(\gamma)\sin(\gamma)C_T(a, \gamma)}{2\beta}\left(1 - \frac{D_r}{\beta x + D_r}\right) \tag{7.14}$$

Howland et al. (2016) also present various studies conducted to assess the centre of wake deflections. Compared to Equation (7.14) most studies show a higher deflection, up to 35% in the extreme case. Therefore it is pragmatically proposed to adjust Equation (7.14) by 25%, i.e.,

$$y_C(x, \gamma) = 1.25D_r\frac{\cos^2(\gamma)\sin(\gamma)C_T(a, \gamma)}{2\beta}\left(1 - \frac{D_r}{\beta x + D_r}\right) \tag{7.15}$$

Figure 7.4 depicts the centre of wake deflections for a yaw angle of $\gamma = 30°, x/D_r = 6$.

No explicit formulas are given for the diameter of the wake at downstream distance x, and therefore we assume that the diameter is given by Equation (7.4) as a starting point. It should be recognized that wind speed across the wake vertically is not uniform. Howland et al. (2016) present results from experimental studies. Figure 7.5 shows the time averaged streamwise ve-

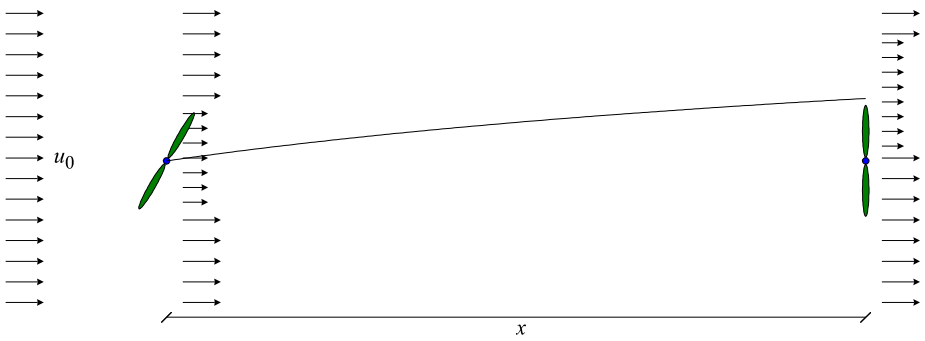


Figure 7.4: Centre of wake deflections for a yaw angle of $\gamma = 30°$, $x/D_r = 6$

locity, $u_w(x)$ for different downstream distances x and distance y from the parallel line following the wind direction from the turbine.

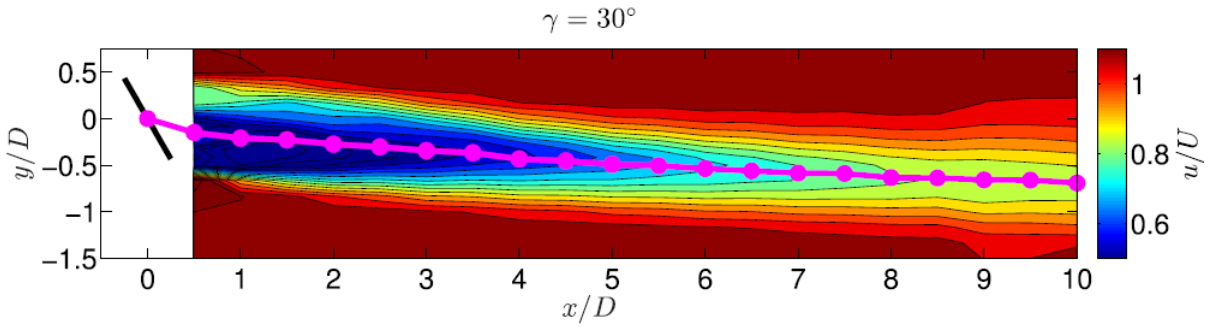


Figure 7.5: Time averaged streamwise velocity, $u = u_w(x)$, contour plot at hub height, taken with a hot-wire probe. The mean velocity is normalized by free-stream velocity $U = u_0 = 12$ m/s. The dark black line represents the yawed turbine. The XY centre of the wake $y_C(x, \gamma = 30°)$ is shown in filled magenta circles (Howland et al., 2016). Note that the definition of the yawing angle is opposite as compared to Figure 7.4.

7.4.1 Wind velocity contours

Figure 7.5 clearly indicates that the velocity reduction is highest at the wake centreline. From Equation (7.12) we may obtain the average speed at distance x downstream the turbine. The diameter of the wake at distance x is given by Equation (7.4), i.e., $D_w = D_r + 2\alpha x$. Let $y = y(x)$ be the horizontal distance from the centreline of the wake. When $y = \pm D_w/2$ the velocity is per definition u_0. Let $u_m = u_m(x)$ be the *minimum* velocity, i.e., at the centreline of the wake. Finally let $u = u(y)$ be the velocity at distance y from the centreline of the wake, where $u(0) = u_m$ and $u(\pm D_w/2) = u_0$. An infinite number of functional forms for $u(y)$ exist and a reasonable form

could be:

$$u(y) = u_{\mathrm{m}} + (u_0 - u_{\mathrm{m}})(2|y|/D_{\mathrm{w}})^c \tag{7.16}$$

where $|x|$ is the absolute value of x and c is a constant. Assuming a circular wake downstream the turbine a mass balance argument requires:

$$\rho u_{\mathrm{w}} \pi (D_{\mathrm{w}}/2)^2 = \rho \int_0^{D_{\mathrm{w}}/2} u(y) 2\pi y dy \tag{7.17}$$

and inserting Equation (7.16) yields:

$$u_{\mathrm{w}}(D_{\mathrm{w}}/2)^2 = \int_0^{D_{\mathrm{w}}/2} \left(u_{\mathrm{m}} y + (u_0 - u_{\mathrm{m}}) 2^{c+1} y^{c+1} / D_{\mathrm{w}}^c\right) dy \tag{7.18}$$

$$= u_{\mathrm{m}} \frac{1}{4} D_{\mathrm{w}}^2 + (u_0 - u_{\mathrm{m}}) \frac{1}{2(c+2)} D_{\mathrm{w}}^{c+1} \tag{7.19}$$

simplifying:

$$u_{\mathrm{w}} = u_{\mathrm{m}} + (u_0 - u_{\mathrm{m}}) \frac{2}{(c+2)} D_{\mathrm{w}}^2 \tag{7.20}$$

and solving wrt. u_{m} gives:

$$u_{\mathrm{m}} = \frac{u_{\mathrm{w}} - u_0 \frac{2}{(c+2)}}{1 - \frac{2}{(c+2)}} \tag{7.21}$$

Assuming $c = 1$, i.e., velocity is increasing linearly from the centreline to the end of the wake, the minimum wake velocity at distance x is given by:

$$u_{\mathrm{m}} = 3u_{\mathrm{w}} - 2u_0 \tag{7.22}$$

It should be noted that Howland et al. (2016) show that the the velocity contours at a given distance downstream the turbine are not perfect circular, so the arguments leading to Equations (7.21) and (7.22) have some limitations.

Example 7.1

We will obtain relevant velocities given the following parameters:

- $\gamma = 30°$
- $D_{\mathrm{r}} = 100$ (m)
- $u_0 = 10$ (m/s)

- $x = 8D_r$

The following Python script implements the relevant functions required for our calculations:

```
'''
alpha = Expansion factor = 0.04
gamma = Yaw angle
x     = Distance downstream the Yawed turbine
u_0   = Free-stream wind velocity
u_w   = Average wind velocity at distance x
D_r   = Diameter of the turbine
u_m   = Minimum speed of the wake at distance x, i.e., at the wake
                                        centreline
c     = Shape factor of the wind velocity profile in the y-direction at
                                        distance x
y     = Distance from the wake centreline
'''
import math
alpha = 0.04
def C_T(gamma): # Trust factor
    return 8*((math.cos(gamma))**2)/9

def u_w(x, u_0, gamma, D_r):
    return u_0 * (1-(1-math.sqrt(1-C_T(gamma))) *(D_r/(D_r+2*alpha*x))**2)

def u_m(u_w, u_0, c = 1):
    f = 2/(2+c)
    return (u_w-f*u_0)/(1-f)

def u_y(x, y, u_0, gamma, D_r, c = 1):
    y_min = u_m(u_w(x, u_0, gamma, D_r), u_0, c)
    y = abs(y)
    if y > (D_r/2 + alpha * x): # Outside the wake
        return u_0
    else:
        return y_min + (u_0 - y_min) * (y/(D_r/2 + alpha * x))**c

def y_C(x, gamma, D_r):
    beta = 2*alpha
    return 1.25*D_r*math.cos(gamma)*math.cos(gamma)*math.sin(gamma)*C_T(
                                        gamma)*(1-D_r/(beta*x+D_r))/(2*
                                        beta)
D_r = 100 # Turbine diameter = hundred metres
x = 8*D_r # Turbine spacing downstream is 8 times the turbine diameter
u_0 = 10  # Free wind velocity = 10 m/2
gamma = 30*math.pi / 180 # Yaw angle = 30 degrees
```

The centreline of the wake at $x = 8D_r$ is found to be $y_C \approx 76$ metres. The velocity at the centreline is found to be $y_m \approx 5.3$ (m/s). The average wake velocity in the $y-z$ plane is found to be $y_w \approx 8.4$ (m/s). Relative to the wake centreline the downstream turbine is positioned a distance $y_C \approx 76$ below this line, hence to find the velocity at the turbine hub we calculate $u_y = u_y(y = -y_C) \approx 9.7$ (m/s). This means that with a yaw angle of $\gamma = 30°$ we have almost avoided the wake effect for the downstream turbine. If we repeat the calculation for $\gamma = 15°$ we obtain a velocity of $u_y = u_y(y = -y_C) \approx 8.4$ (m/s) which is a significant reduction compared to $u_0 = 10$ (m/s). □

7.4.2 Turbulence intensity

In addition to the reduced velocity caused by the wake there is also a significant negative turbulence impact of the wake. The turbulence intensity is usually measured in terms of the variation of the velocity, or more precisely as the Root-Mean-Square (RMS)of the turbulent velocity fluctuations at a particular location over a specified period of time.

Experimental studies by Howland et al. (2016) indicate that the turbulence intensity contours in the $y-z$ plane are similar to the velocity contours. Although these are not circular, to simplify we propose to model the contours as circles. The standardised turbulence intensity, TI shows a maximum intensity around $\text{TI}_{\text{mx}} = 12$ at distance $x = 8D_r$ in the study by Howland et al. (2016). Again we propose an intuitive function for the turbulence intensity relative to the wake centreline:

$$\text{TI}(y) = \text{TI}_{\text{mx}} \left[1 - \left(\frac{2|y|}{D_w} \right)^d \right] \tag{7.23}$$

where $D_w = D_r + 2\alpha x$. d is a shape parameter to be chosen. The simplest form of the turbulence intensity function is achieved for $d = 1$ which is the recommended default value.

Note that TI_{mx} in Equation (7.23) has not been specified. If we can measure the turbulence intensity under various operational conditions in the centerline of the wake, these measurements can be used in Equation (7.23). If such information is not available it is reasonable to believe that the turbulence intensity is proportional to the time average centreline velocity given by Equation (7.21). This will usually be sufficient for our modelling. The reason for such an argument is that to explicit consider the negative impact of turbulence, we would link the turbulence intensity to the degradation rate of the system or item analysed. For example if we represent degradation by a Wiener process with drift, i.e., $\{X_t = \mu t + \sigma W_t, t > 0\}$, the objective is to link the drift parameter μ to the turbulence intensity. Assuming a linear relationship we could

then define the following function for the drift parameter:

$$\mu = \mu(\mathrm{TI}) = \mu_0 + \frac{(\mu_{\mathrm{mx}} - \mu_0)\mathrm{TI}}{\mathrm{TI}_{\mathrm{mx}}} \tag{7.24}$$

where μ_0 is the drift parameter (degradation speed) without negative turbulence impact, and μ_{mx} is the drift parameter if the turbine is placed in the centreline of the wake.

Example 7.2

We consider two turbines as shown in Figure 7.4. The turbine to the left is denoted turbine 1 and the downstream turbine to the right is denoted turbine 2. To simplify we assume a constant free wind velocity u_0, and the wind direction is parallel to the line connecting turbine 1 and turbine 2. The challenge is to optimize the yawing angle. To simplify we assume that degradation of Turbine 1 is not affected by the yawing angle, hence we do not include maintenance and failure cost related to turbine 1. For turbine 2 we consider maintenance cost and we apply the cost model in Section 5.6.1. The following Python script shows the model parameters:

```
u_0   = 10 # Free wind velocity [m/s]
rho   = 1.225 # Air density at 15 degrees [kg/m^3]
r     = 60 # Rotor radius [m]
D_r   = 2*r # Rotor diameter [m]
alpha = 0.04 # Wake decay constant
x     = 7*2*r # Distance between the turbines
MTTF  = 8760*5 # Mean time to failure without maintenance = 5 years
l     = 100 # Failure limit, normalized to 100%
mu_0  = 1/MTTF # Drift parameter, no wake effect
mu_x  = 5*mu_0 # Higher degradation at wake centreline
sigma = 50*mu_0 # Infinitesimal standard deviation, volatility
T_l   = 2*7*24 # Lead time of renewal/maintenance = 2 weeks
c_R   = 2500000 # Cost of renewal
c_F   = 7500000 # Cost of failure
p_e   = 0.5 # Energy price (NOK/kWh)
```

To optimize the yaw angle we essentially calculate the following:

- Profit per unit time for turbine 1
- Profit per unit time for turbine 2, assuming no failures
- $C(m)$ whis is renewal, failure and downtime cost per unit time according to Equation (5.15)

These cost per unit time can be calculated for each value of the yawing angle γ, where $C(m)$ also is minimized wrt the maintenance limit m. As the yawing angle increases the reduction factor $C_{\mathrm{P}}(\gamma)$ decreases and hence the profit. Further an increased yawing angle moves the wake

away from turbine 2. This means that the wind velocity increases yielding higher profit, and the turbulence intensity decreases giving less degradation on turbine 2.

Figure 7.6 shows the result from the calculations. The optimal yawing is found to be 24°. Note as the yawing angle approaches 30° the distance to the centre of the wake is starting to decline, hence there is no reason to increase the yawing angle any more.

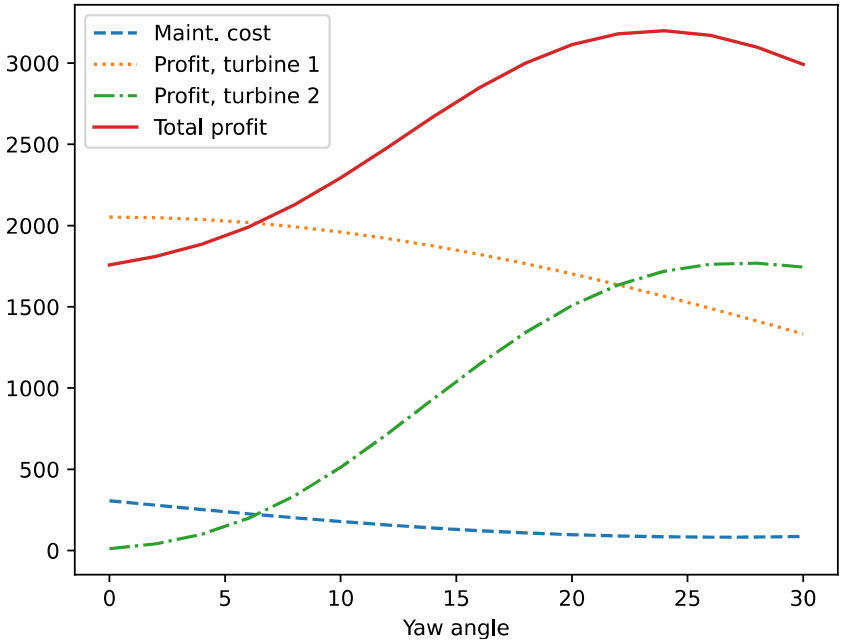


Figure 7.6: Profit for turbine 1 and 2, maint. and downtime cost for turbine 2, and total profit

□

Problems

7.1 Consider Example 7.1. Calculate u_{m} for $\gamma = 5, 10, \ldots, 30$ for $x = 5D_{\mathrm{r}}$ and $x = 8D_{\mathrm{r}}$ respectively.

7.2 Consider Example 7.2. Implement the required Python code to produce the results shown in Figure 7.6.

Chapter 8

Grouping and opportunity maintenance

8.1 Introduction

In classical maintenance optimization the objective is to find the optimum frequency of maintenance of one component at a time. However, in the multi-component situation there exist dependencies between the components, e.g., they may share a common set-up cost (economy of scope), the costs may be reduced if the contract to a maintenance contractor is huge (economy of scale), etc.

In this presentation we will introduce some rather simple approaches for maintenance grouping and opportunity maintenance. We basically consider the following cost elements:

- Man-hour costs and material costs related to preventive maintenance of each component
- Set-up costs to get access to the components to be maintained, and by paying the set-up costs access to several components is obtained. We limit the scope to consider a one level structure of set-up costs, meaning that the set-up cost is the same for all components. In a multi level structure the set-up cost could be split into a general set-up cost for accessing e.g., a location/cite, and further into set-up cost for a group of components related to e.g., preparing of the work for these components.
- Costs of taking a component out of service. These costs are included in the set-up costs from a modelling point of view.
- Man-hour costs and material costs related to corrective maintenance. Typically set-up costs can not be shared by other components unless preventive maintenance is advanced (opportunity maintenance).
- Costs related to the effect of a failure, i.e., punctuality, unavailability, safety and material damage costs

We often distinguish between the static and the dynamic planning regimes. In the static regime the grouping is fixed during the entire system lifetime, whereas in the dynamic regime the groups are re-established over and over again. The static grouping situation may be easier to implement than the dynamic, and the maintenance effort is constant, or at least predictable. The advantage of the dynamic grouping is that new information, unforeseen events, etc., may require a new grouping and changing of plans.

The presentation here discusses how we can formalize the optimization of maintenance grouping. I.e., we seek to group maintenance tasks so that total costs are minimized. To summarize the difference between static and dynamic grouping we have:

- Static grouping where the groups are fixed
  - It is always the same maintenance task that are included in the same group, and each maintenance group is performed at a fixed interval.
  - In the work order system one group is specified as one work order repeating every τ_i unit of time.
- Dynamic grouping where we create the groups "on the fly"
  - The time of next maintenance is recalculated in principle continuously
  - The set of maintenance tasks going into a group is varying form time to time
  - In principle we can plan for several groups ahead, but often we only consider the first group of tasks
  - We can update the plan if we get new information, or there are additional opportunities to carry out maintenance
  - The downside is significantly more administrative work and challenges in relation to staff planning.

For an introduction to maintenance grouping we refer to Wildeman (1996) who discusses these different regimes in detail.

Maintenance tasks are here preventive tasks where the base interval is calendar controlled or controlled by runtime. At the end of the presentation we also discuss condition-based maintenance.

The costs of disassembling and re-assembling are here included in the set-up cost. In the model presented we also assume that the set-up costs are the same for all activities. It is further assumed that there is one and only one maintenance activity related to each component. This simplifies notation because we then may alternate between failure of component i and executing maintenance activity i where there is a unique relation between component and activity.

The basic notation to be used is below. The terms maintenance task and maintenance activity are used interchangeably. Table 8.1 shows the notation used. Note that t is used to represent calendar/global time or accumulated mileage for a car or a train. x is used to represent local time, i.e., time since last maintenance.

Table 8.1: Notation used in grouping

c_i^{P}	Planned maintenance cost, exclusive set-up cost for activity i. Typically the costs of replacing one unit periodically
c_i^{U}	Unplanned costs upon a failure of component i. These costs include the corrective maintenance costs, safety costs, punctuality costs, unavailability costs and costs due to material damage.
S	Set-up cost, i.e., costs for preparations, access etc which can be “shared" by several PM activities
$\lambda_{\mathrm{E},i}(x)$	Effective failure rate for component i. Here the argument x represents local time since last maintenance
$M_i(x)$	$= xc_i^{\mathrm{U}}\lambda_{\mathrm{E},i}(x)$ = Accumulated expected costs due to failures in a period $[0,x)$ for component i maintained at time 0, exclusive planned maintenance cost
$C_i(x,k)$	$= [c_i^{\mathrm{P}} + S/k + M_i(x)]/x$ = Expected total costs per unit of time for component i for a maintenance cycle of length x if setup costs are shared by k activities
$x_{i,k}^*$	Maintenance interval that minimizes $\Phi_i(x,k)$ if setup costs are shared by k activities
$C_{i,k}^*$	Minimum cost for a component i maintained at optimal interval
k_i	Average number of components sharing the set-up costs for the i'th component, i.e., the i-th component is in average maintained together with $k_i - 1$ other components
C_i^*	Average minimum costs per unit time over all k-values
x_i^*	Optimum value of x_i over all k-values. x_i^* is measured since last maintenance on component i
t_0	Point of time when we are planning the next group of activities. Initially $t_0 = 0$. t_0 is measured in running time since $t = 0$.
x_i	Age of component i at time t_0, i.e., time since last preventive maintenance activity
t_i^*	$t_{i,}^* = t_0 + x_i^* - x_i$ = optimum time for execution in the average situation
$G(g)$	Candidate group, i.e., the set of the first g components to be maintained according to individual schedule with $t_{i,\mathrm{Av}}^*$ as the basis for due time
l_i	How often a component is utilizing the maintenance opportunity in static indirect grouping
N	Number of activities/components
T	For dynamic grouping T is the end of planning horizon, i.e., we are planning from $t_0 = 0$ to T. For indirect static grouping T is used as the lowest interval.
T_j	Interval for group j in static direct grouping.

8.2 Static grouping

For static grouping, we distinguish between indirect and direct grouping:

- Indirect grouping means that the groups are not established by a direct rule, but that the groups are established based on a principle. This principle is that an activity can be executed on each maintenance opportunity, every second opportunity and so on. How often to be executed is then the optimization challenge.
- Direct grouping means that the groups are established by investigating the intervals one by one and form groups of activities activities having approximately the same interval.

8.2.1 Indirect static grouping

The indirect grouping principle is that the time of each activity is determined indirectly by specifying how often the task is performed relative to a fixed repetitive time of maintenance. The situation now is as follows:

- There is a possibility to do preventive maintenance at point of times $T, 2T, 3T, \ldots$
- For component i this opportunity is utilized every l_i'th time, i.e., the interval between maintenance for this component is $l_i T$
- The challenge is to determine T and $l_i, i = 1, 2, \ldots$

For a given value of T and $l_1, l_2, \ldots$ the expected cost per unit time is:

$$\begin{aligned} C(T, l_1, l_2, \ldots) &= S/T + \sum_{i=1}^{n} \left[c_i^{\mathrm{P}} + M_i(l_i T) \right] / (l_i T) \\ &= S/T + \sum_{i=1}^{n} \left[c_i^{\mathrm{P}} / (l_i T) + c_i^{\mathrm{U}} \lambda_{\mathrm{E},i}(l_i T) \right] \end{aligned} \tag{8.1}$$

where $M_i(x)$ is the total failure related cost in a period of length x since last maintenance.

8.2.2 Heuristic for indirect static grouping

Minimizing Equation (8.1) is a mixed-integer optimization problem. Generally such problems need to be solved by heuristics when N becomes large. The following heuristic is suggested to find a reasonably good solution:

1. For each activity i we find the value of τ_i which minimizes $C(\tau_i) = \left(S + c_i^{\mathrm{P}}\right) / \tau_i + c_i^{\mathrm{U}} \lambda_{\mathrm{E},i}(\tau_i)$
2. An initial value of T is set equal to the lowest value of the τ_i-values
3. Chose $l_i \approx \tau_i / T$ (nearest integer)
4. Keep the l_i's fixed, and minimize $C(T, l_1, l_2, \ldots)$ with respect to T

5. GoTo 3 and change the l_i's ± 1 one by one to search for better solutions

6. An approximate optimal solution is found when the iteration scheme does not improve the solution, i.e., we do not find a solution with a lower expected cost

8.2.3 Direct static grouping

By direct grouping, the groups are selected directly by inspecting individual intervals. Tasks are now split into m non-overlapping groups, $G_1, G_2, \ldots$. Activities in a group are maintained at the same time. The groups are established so that activities in a group have approximately equal intervals. For group j, we let T_j denote the interval for this group. Total expected costs per unit time is given by

$$C(T_1, T_2, T_3, \ldots, T_m) = \sum_{j=1}^{m} \left(S/T_j + \sum_{i \in G_j} \left[c_i^{\mathrm{P}} / T_j + c_i^{\mathrm{U}} \lambda_{\mathrm{E},i}(T_j) \right] \right) \tag{8.2}$$

Heuristic for direct static grouping

The following heuristic is proposed for obtaining a reasonable good solution:

1. For each activity find the value τ_i which minimizes $C(\tau_i) = (S + c_i^{\mathrm{P}})/\tau_i + c_i^{\mathrm{U}} \lambda_{\mathrm{E},i}(\tau_i)$

2. Sort in increasing order, i.e., $\tau_{(1)} \leq \tau_{(2)} \leq \ldots$

3. Look for natural clusters in the intervals, and let these forms groups $G_1, G_2, \ldots$

4. Given a split into groups, i.e., $G_j, j = 1, 2, \ldots, m$, minimize the cost Equation (8.2) with respect to $T_1, T_2, \ldots, T_m$

5. GoTo 3 and varying the groups to look for better solutions, for example moving one activity from one group to another group, merge two groups, or split one group into two groups

6. An approximate optimal solution is found when the iteration scheme does not improve the solution.

8.3 Dynamic grouping

In dynamic grouping there are no fixed group. At a given point of time, t_0, we start forming the next group based on "individual" due dates. Figure 8.1 illustrates the situation for four components maintained at t_1, t_2, t_3 and t_4 in the past, where the due dates t_1^*, t_2^*, t_3^* and t_4^* are based on the individual optimal intervals x_1^*, x_2^*, x_3^* and x_4^*.

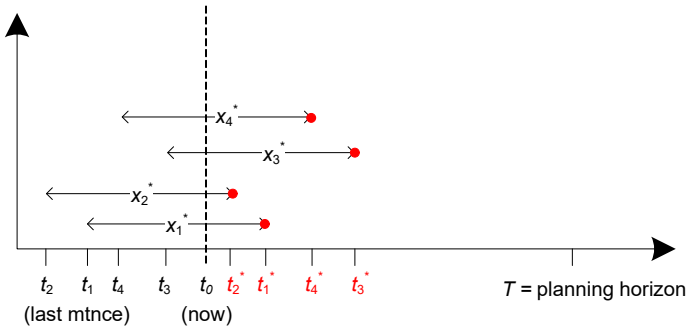


Figure 8.1: Grouping - Due dates. The t_i's are the point of time of last maintenance, the x_i^*'s are the individual optimal intervals, and the t_i^*'s are the corresponding due dates.

In the optimization we consider the time from now on, t_0, up to the end of planning horizon, T. Given the information we have at time t_0 we can form the next group and when to execute the corresponding maintenance activities. We can also form the second group, the third group etc. But then we should realize that we might get new information later on, and hence have to reschedule some future groups.

In this situation there is no single cost equation to optimize. We will structure the cost elements and then propose a heuristic for forming groups.

For each component i there is an expected time dependent cost which is a function of the time since the last preventive maintenance activity, i.e., $M_i(x)$. In order to establish $M_i(x)$ we need (i) to establish the accumulated expected number of failures in the period $[0,x)$, (ii) we need to specify the expected corrective maintenance costs for the repair of each failure, and (iii) we have to specify the impact of the failure on safety, production, etc., and quantify these into cost figures. In the model presented here we assume that the effective failure rate, $\lambda_{\mathrm{E}[,]i}(x)$ may be established for the different failure characteristic, and maintenance strategies (e.g., periodic replacement and condition monitoring). Next the costs associated with a failure of component i may in principle be found by risk modelling, reliability modelling. The result of such modelling is one figure for the expected unplanned cost of failure, i.e., c_i^{U}. We have

$$M_i(x) = x c_i^{\mathrm{U}} \lambda_{\mathrm{E},i}(x) \tag{8.3}$$

The planned costs comprise the costs of executing the maintenance on component i (c_i^{P}) and set-up costs (S) of getting access to the component. The set-up costs may in general be shared with $k-1$ other activities. The average contribution to the total costs for component i per unit time is given by:

$$C_i(x,k) = \left[c_i^{\mathrm{P}} + S/k + M_i(x)\right]/x \tag{8.4}$$

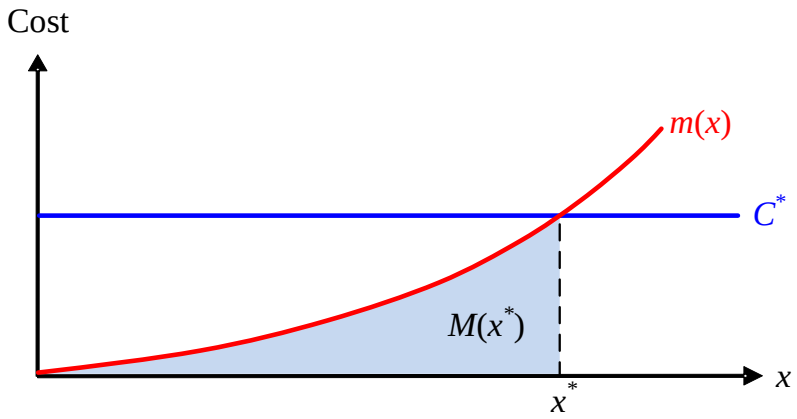


Figure 8.2: Marginal cost consideration

If the grouping was fixed, i.e. static grouping, the optimization problem would just be to minimize $C_i(x,k)$ wrt x for all k components maintained at the same time. For dynamic grouping the mathematical challenge is now to establish the grouping either in a finite or infinite time horizon. In addition to the grouping, we also have to schedule the execution time for each group (maintenance package). The grouping and the scheduling can not be done separately. Generally, such optimization problems are NP hard (see Garey and Johnson, 1977, for a definition), and heuristics are required. Before we propose our heuristic we present some motivating results.

Let $C^*_{i,k}$ be the minimum average costs when one component is considered individually, and let $x^*_{i,k}$ be the corresponding optimum x value. It is rather easy to prove that

$$m_i(x^*_{i,k}) = M'_i(x^*_{i,k}) = C^*_{i,k} \tag{8.5}$$

meaning that when the instantaneous expected unplanned costs per unit time, $m_i(x)$, exceeds the average costs per unit time, maintenance should be carried out. The way to use the result is now the following. Assume we are going to determine the first point of time to execute the maintenance, i.e., to find t^*_{i,k_i} starting at $t = 0$. Further, assume that we know the average costs per unit time, C^*_{i,k_i} but that we have for some reason "lost" or "forgotten" the value of the optimal interval, x^*_{i,k_i}. What we then can do is to find t such that $m_i(t) = M'_i(t) = C^*_{i,k_i}$ yielding the first point of time for maintenance, see Figure 8.2 for an illustration. Then from time t and the remaining planning horizon we can pay $C^*_{i,k}$ as the minimum average costs per unit time. This is the traditional *marginal costs* approach to the problem, and brings the same result as minimizing Equation (8.4).

The advantage of the marginal thinking is that we now are able to cope with the dynamic grouping. Assume that the current time is t_0, and x_i is the age (time since last maintenance) for component i in the group we are considering for the next execution of maintenance. Further assume that the planning horizon is $[t_0, T)$. The problem now is to determine the point of time $t(\geq t_0)$ when the next maintenance is to be executed. The total costs of executing the maintenance activities in a group is

$$C_{\mathrm{P}} = S + \sum_i c_i^{\mathrm{P}} \tag{8.6}$$

which we pay at time t. Further, the expected unplanned costs in the period $[t_0, T)$ is

$$C_{\mathrm{U}} = \sum_i \left[M_i(t - t_0 + x_i) - M_i(x_i) \right] \tag{8.7}$$

where x_i is the (local) age of component i at time t_0. Note that $M_i(t - t_0 + x_i)$ is the expected cost from the last maintenance of component i until it will be preventively maintained at time t. From this value we subtract the expected cost $M_i(x_i)$ already "paid" at time t_0. Note that at time t_0 we know the "history" of component i since the last maintenance, i.e., x_i time units ago. We might use this information to get a more correct expression for the expected cost in the interval $[t_0, t)$. It is not always easy to obtain such an expression, hence we often approximate with Equation (8.7).

For the remaining time of the planning horizon the total costs are

$$C_{\infty} = (T - t) \sum_i C^*_{i,k_i} \tag{8.8}$$

provided that each component i can be maintained at "perfect match" with $k_i - 1$ activities the rest of the period. Since $C^*_{i,k}$ depends on how many components that share the set-up cost, which we do not know at this time, we use some average value Φ^*_i. We assume that we know this average value at the first planning. To determine the point of time for maintaining a given group of components, say $G(g)$ with the g first activities we thus minimize:

$$c_1(t; g) = S + \sum_{i \in G(g)} \left[c_i^{\mathrm{P}} + M_i(t - t_0 + x_i) - M_i(x_i) + (T - t) C^* \right] \tag{8.9}$$

The costs in Equation (8.9) depend on which components to include in the group of activities to be executed next. The more activities we include, the higher the costs will be. For some activities it might thus be cheaper to include them in groups to be executed later. For activities we do not include in this first group we assume that they will be maintained at their "optimum"

time $t_i^*, > t$. The total contribution to the costs related to these activities in $[t_0, T)$ is:

$$c_2(t;g) = \sum_{i \notin G(g)} \left[c_i^{\mathrm{P}} + S/k_i + M_i(x_i^*) - M_i(x_i) + (T - t_i^*) C_i^* \right] \tag{8.10}$$

provided they can be maintained at "perfect match" with other activities, i.e., the set-up costs are shared with $k_i - 1$ activities, and executed at time t_i^*. The total optimization problem related to the next group of activities is therefore to minimize:

$$\begin{aligned} c(t;g) = S + & \sum_{i \in G(g)} \left[c_i^{\mathrm{P}} + M_i(t - t_0 + x_i) - M_i(x_i) + (T - t) C_i^* \right] \\ & + \sum_{i \notin G(g)} \left[c_i^{\mathrm{P}} + S/k_i + M_i(x_i^*) - M_i(x_i) + (T - t_i^*) \Phi_i^* \right] \end{aligned} \tag{8.11}$$

The idea is simple, we first determine the best group to execute next, and the best time to execute it. Further we assume that subsequent activities can be executed at their individual optimum. It is expected to do better by taking the second grouping into account when planning the first group, and not only treat the activities individually. See e.g., Buday et al (2005) for more advanced heuristics in similar situations to those presented here. The heuristic is as follows:

Step 0 - Initialization

This means to find initial values of the k_i's and use these k-values as basis for minimization of Equation (8.11). This will give initial values for the x_i^*'s and the corresponding C_i^*'s. Finally the time horizon for the scheduling is specified, i.e., we set $t_0 = 0$ and choose an appropriate end of the planning horizon (T).

Step 1 - Prepare for defining the group of activities to execute next

Calculate tentative due dates $t_i^* = (x_i^* - x_i) + t_0$ for all activities, and sort in increasing order. See Figure 8.1 for an illustration.

Step 2 – Establish the candidate groups

For $g = 1, 2, \ldots, N$ we use the ordered t_i^*'s to find a candidate group $G(g)$ of size g to be executed next. If $t_g^* > \min_{i<g}(t_i^* + x_i^*)$ this means that at least one activity in the candidate group needs to be executed twice before activity g is scheduled which does not make sense. Hence, in this situation the last candidate group, $G(g)$ is dropped and we are not searching for more candidate groups at the time being.

Step 3 - Find optimum execution time for each candidate group

For each candidate group $G(g), g = 1,2,\ldots$, minimize $c(t,g)$ in Equation (8.11) with respect to execution time t. Next choose the candidate group $G(g)$ that gives the minimum cost. This group should then be executed at the corresponding optimum time t.

Step 4 – Prepare for subsequent groups

We assume that all activities in the chosen candidate group are executed at time t. This corresponds to setting $x_i = 0$ for $i \in G(g)$, $x_i = x_i + t - t_0$ for $i \notin G(g)$ and then update the current time, i.e., $t_0 = t$. If $t_0 < T$ GoTo Step 1, else we are done.

There are several ways to improve the algorithm. One intuitive improvement is to improve the estimates of k_i and corresponding x_i^* and Φ_i^* to be specified in Step 0. This is easy, since we in Step 4 get a new value of k for those activities included in the candidate group, and when the algorithm terminates we simply set k_i as the average for each activity i in the period $[0,T)$. We may then start over again at Step 0 with these new values of k_i.

8.4 Opportunity based maintenance

The dynamic scheduling regime presented above is a good basis for opportunity based maintenance. The scheduling we have proposed may be used to set up an explicit maintenance plan for the time horizon $[0,T)$. But even though the plan exists, we may consider changing it as new information becomes available, either in terms of new reliability parameter estimates, or if unforeseen failures occur. In operation, for any time t_0 we may update the scheduling of preventive maintenance.

Now assume that the due date t for the next scheduled maintenance of group $G(g)$ is larger than the current time t_0. Assume that a failure has occurred or there is another event occurring at time t_0 giving an opportunity to save the setup-cost S if we execute some preventive maintenance activities. Some, or all of the activities in $G(g)$ should now be considered for execution given the opportunity at time t_0. If $t_i^* \le t_0, i \le g$ this means that these activities have individual due dates in the past, hence it is obvious that these activities should be executed at this given opportunity.

Activities not scheduled in $G(g)$ should not be executed since they were not even included in a group to be executed later than t_0. The basic question is thus which of the remaining activities in $G(g)$ should be executed. Assume that we have found that it is favourable to execute the first $i-1 < g$ activities on this opportunity. The procedure to test whether or not activity i also should be executed is as follows:

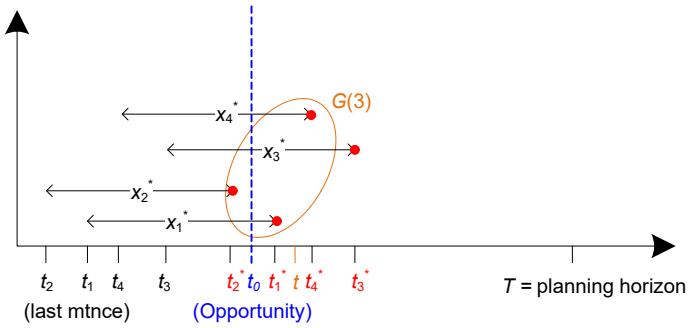


Figure 8.3: Opportunity maintenance. There is an opportunity at time t_0, where group $G(3)$ was scheduled for execution at time t.

- First we assume that all activities up to i are executed on this opportunity , i.e., we set $x_j = 0, j \leq i$, and for activities above activity i, i.e., $j > i$, we set $x_j = t_0 - t_j$, and will evaluate the next group to be executed at some time $t' \geq t$:
- Let $C_1 = c_i^{\mathrm{P}} + \min_{t',g'} c(t',g')$, i.e., the best we can do if we decide to execute activity i at time t_0
- Next, we assume that only activities up to $i-1$ are executed at time t_0, i.e., $x_j = 0, j \leq i-1$, and $x_j = t_0 - t_j, j \geq i$, where we also evaluate the next group:
- Let $C_2 = \min_{t',g'} c(t',g')$, i.e., the best we can do if we decide not to execute activity i at time t_0
- If $C_1 > C_2$ is it not beneficial to do activity i.

If it was beneficial to do activity i at t_0 we should test for $i = i+1$ and repeat as long as $i \leq g$.

Figure 8.3 illustrates the situation. Assume that at the time of the opportunity we had already scheduled $G(3)$ for execution at time t. The individual due date for activity 2 has been passed, hence activity 2 will be executed. Then we consider if it pays off to execute also activity 1. If so, typically activity 3 and 4 will be a new group, say $G(2)$ to be executed at some time $t' > t$.

Problems

8.1 Consider a situation where we have 4 components. We will establish a standard indirect static grouping strategy. The following data is provided: $S = 2, c_1^P = 2, c_2^P = 1, c_3^P = 3, c_4^P = 1. c_1^U = 5, c_2^U = 50, c_3^U = s_4^U = 10. \mathrm{MTTF}_1 = 4, \mathrm{MTTF}_2 = 3, \mathrm{MTTF}_3 = 3, \mathrm{MTTF}_4 = 5, \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 3$.

Find tentative optimal intervals for each component if they are maintained individually and where we assume that we do not have to pay the set-up cost. Use this to find tentative values for l_i and T. Then try some iterations to see if a better solution can be found. Note, in the heuristic

we proposed to find the individual solutions assuming we pay the entire set-up cost. In this exercise a slightly different approach was proposed, i.e., that we do not have to pay the set-up cost. The two approaches should converge to the same result.

8.2 Consider the situation in Problem 8.1. Repeat the analysis if you in the initiation assume that we have to pay the set-up cost.

8.3 Consider the situation in Problem 8.2. Apply the method of direct static grouping to the same data.

Chapter 9

Spare part optimization

9.1 Introduction

When optimizing models for individual components in relation to the interval τ, it may be appropriate to consider whether it pays to have a spare part in stock allowing us to reduce downtime. In the analysis, we can then compare the situation with and without spare part in stock, and find out if the cost of inventory can be justified.

In many situations there are several components that "fight" for the same spare part, and it becomes a question of *how many* spare parts we need. We can compare this with the situation at home where the question is how many light bulbs (of a given type) we will normally have in stock to avoid not running out of light bulbs. In this lesson, we'll look at two different ways to model this:

- An analytical model where we can set up equations to calculate the expected share of the time we lack one, two or more spare parts
- A Markov model where we can find the same answer, but where we have more flexibility to give in different assumptions

9.2 An analytical model

- Constant failure rate (i.e., the total failure rate for many components that need a new spare part in the event of failure) = λ
- Number of spare parts = s
- The spare parts are stored in a stock and are retrieved from there if necessary
- Failed components are repaired in a workshop

- The number of components under repair in the workshop = X
- Repair rate for each component repaired = μ
- We have endless number of repairmen, i.e., a repairman can always start repairing a component that comes to the workshop

Note that we have assumed that components that fail will be repaired. If we instead have to buy new components in the event of a failure, the model will be identical if we allow the expected time it takes to obtain a new component = $1/\mu$.

9.2.1 Mathematical model

- According to Palm's theorem, $X \sim Po(\lambda/\mu)$
- From the Poisson distribution it follows by introducing $p(k) = \Pr(X = k) = \frac{(\lambda/\mu)^k}{k!} e^{-\lambda/\mu}$:
  - $p(0) = e^{-\lambda/\mu}$
  - $p(s+1) = \frac{\lambda/\mu}{s+1} p(s)$
- The probability of missing spare parts is: $R(s) = \Pr(X > s) = \sum_{k=s+1}^{\infty} p(k)$, which gives:
  - $R(0) = 1 - p(0)$
  - $R(s+1) = \sum_{k=s+2}^{\infty} p(k) = \sum_{k=s+1}^{\infty} p(k) - p(s+1) = R(s) - p(s+1)$
- The number of units we may lack is referred to as BO (Backorders):
  - $\mathrm{EBO}(s) = \mathrm{E}(\mathrm{BO}) = \mathrm{E}(max(0, X-s)) = \sum_{k=s+1}^{\infty} (k-s)p(k)$
  - $\mathrm{EBO}(s+1) = \sum_{k=s+2}^{\infty} (k-s-1)p(k) = \sum_{k=s+1}^{\infty} (k-s-1)p(k)$
  - $\mathrm{EBO}(s+1) = \mathrm{EBO}(s) + \sum_{k=s+1}^{\infty} (-1)p(k) = \mathrm{EBO}(s) - R(s)$
- The following recursive regime can then be used
  - $p(0) = e^{-\lambda/\mu}$
  - $R(0) = 1 - p(0)$
  - $\mathrm{EBO}(0) = \mathrm{E}(X) = \lambda/\mu$
  - $p(s+1) = \frac{\lambda/\mu}{s+1} p(s)$
  - $R(s+1) = R(s) - p(s+1)$
  - $\mathrm{EBO}(s+1) = \mathrm{EBO}(s) - R(s)$

9.2.2 Simple cost model

- Cost elements
  - c_U = Unavailability cost per unit of time
  - c_S = Capital cost per unit of time to keep a unit in stock
- Cost equation, i.e., the objective function:
  - $C(s) = c_S s + c_U \text{EBO}(s)$

To minimize the cost equation, $C(s)$ is calculated for different values of s. We must then use the recursive formulas to find $\text{EBO}(s)$

9.3 Markov modelling

Markov modelling is a special way to model transitions between system states. Here we will investigate Markov models where we have a limited number of states. Each state is given a number (identifier). It turns out appropriate to let the identifier of a state be the number of spare parts in stock. If the stock is empty and no one is requesting a component, we give the state number 0, while negative state numbers correspond to the number of spare parts we have shortages, i.e., a stock-out situation.

Markov models can in some cases be solved analytically, but we usually need a computer program to calculate the Markov models.

The following assumptions and limitations apply:

- Failures and repair times are exponentially distributed
- We can introduce different strategies, e.g., vary how many repair men we want
- For non-exponential repair times, we can use so-called phase type distributions. This is a little more to elaborate, but can provide reasonably good solutions with not too much extra modelling work
- Disadvantages
  - In principle, we may have infinite number of backorders, while in the model we must limit the number of states in the transition matrix, limiting the number of backorders the model can hold
  - We must manually specify the transition matrix, which can be tedious when testing different strategies, with programming this is not that difficult
  - For very large systems, there may be challenges with computational speed

9.3.1 Model specification

- Constant failure rate = λ, i.e., demand rate of spare parts
- Number of spare parts = s
- The spare parts are stored in a stock and are retrieved from there if necessary
- Failed components are repaired in a workshop
- The number of components under repair in the workshop = X
- Repair rate for each component being repaired = μ
- We have a limited number of repair men, and the number = m

Graphical representation

The following are transitions between states. We assume in the first place that we have a large number of repairmen.

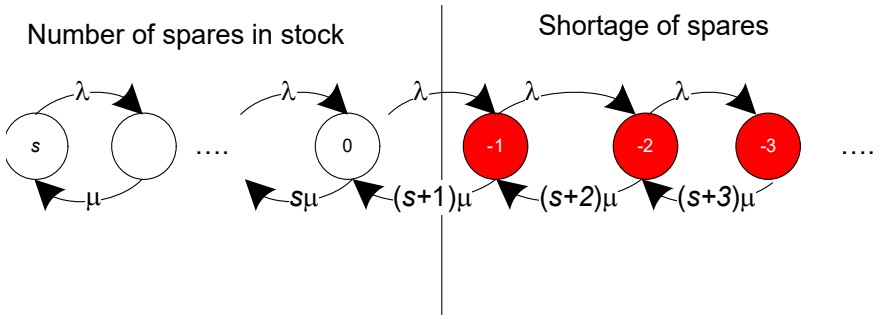


Figure 9.1: Markov transition diagram

From the Markov model we find the steady state solution, and unavailability (or expected backorder) is given by

$$U(s) = \text{EBO}(s) = P_{-1} + 2P_{-2} + 3P_{-3} + ... \tag{9.1}$$

where P_{-1} is the element in the solution vector representing shortage of exactly one item, P_{-2} shortage of exactly two items etc.

9.3.2 m-Repairmen

In the model so far, we have assumed that we have an infinite number of repair men. That means that the more devices that are for repair, the greater the repair rate will be. In general, we have

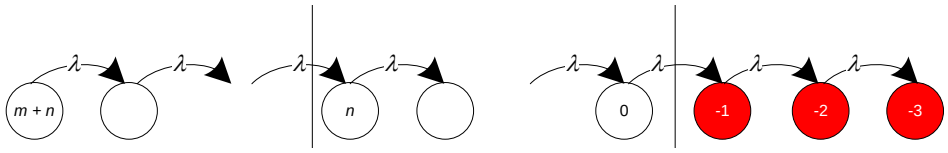


Figure 9.2: Step 1

assumed that the repair rate is $X\mu$, where X is the number of units for repair in the workshop, while μ is the repair rate (completion rate) each repairman has.

If we only have m repairmen, the transition rate is $min(X,m)\mu$, where X is most easily determined by assessing how many units are for repair for the current state. For example, if $s = 5$, and we consider state -1, X will be $5+1 = 6$, and if we have only $m = 4$ repairers, the rate from state -1 to state 0 will be equal to 4μ.

The cost equation is given by:

$$C(s,m) = c_S s + c_U \text{EBO}(s) + c_M m \tag{9.2}$$

where Cc_M is the cost per unit time of having one repairman available. Note that a repair man is doing other tasks, so C_M is not necessarily very large.

9.3.3 A reorder policy model

The following model is similar to the lot size, reorder point policy, (r,Q), used in inventory management. The model assumptions are:

- Constant rate of failure λ
- Mean lead time when ordering new spares = MLT
- Lead times are Gamma (Erlang) distributed with parameters $\alpha = 4$, and $\mu = \alpha/\text{MLT}$
- Totally m new spares are ordered when stock level equals n

Note that $\alpha = 4$ may be changed to account for general value of SD(LT) = $\alpha^{1/2}/\mu$

Figure 9.2 illustrates the situation for taking components out of the stock. Initially we have $m+n$ components in stock. Then when the level reaches n, i.e., the re-order point, an order is placed for replenishment of the stock.

To model the lead time, we introduce intermediate states representing the gamma distribution, i.e., a transition from state n to state n_1 to state n_2 to state n_3 and finally to state $m+n$. Figure 9.3 illustrates this.

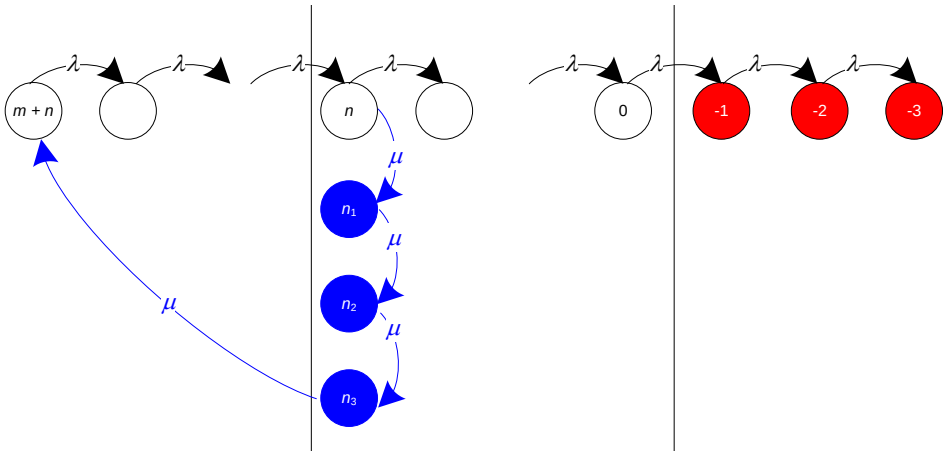


Figure 9.3: Step 2

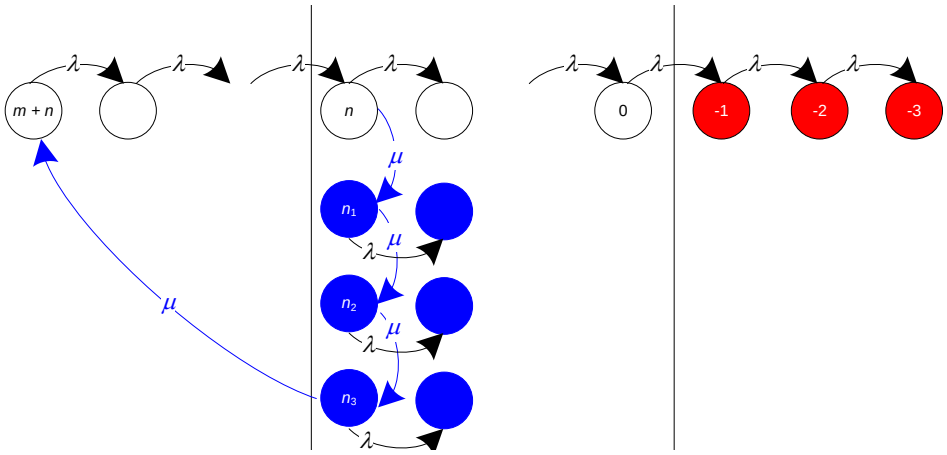


Figure 9.4: Step 3

During the lead time, there might be a new demand for a spare (i.e., a failure). This means that stock level is reduced by one. Figure 9.4 illustrates this by the transitions from n_1 to $(n-1)_1$ and so on.

Figure 9.5 illustrates the complete picture.

Optimization

To optimize the model, we need to specify

- c_F = Fixed cost per order
- c_H = Holding cost per item per unit time

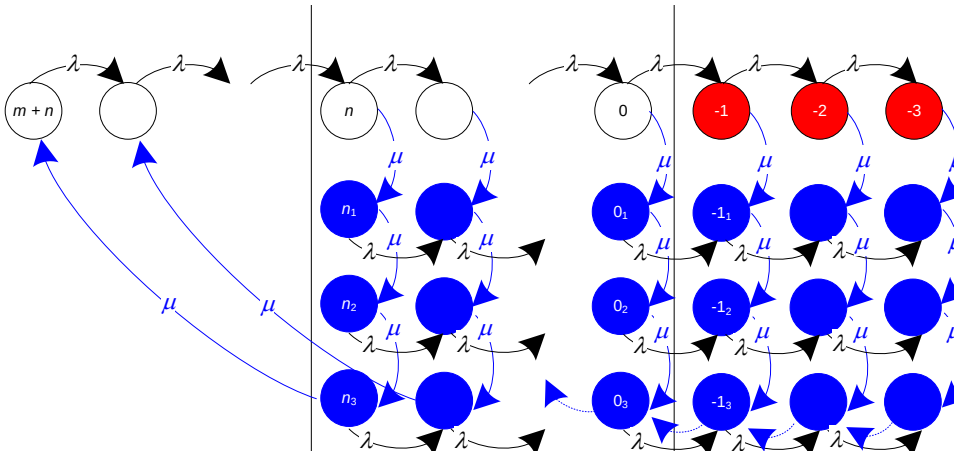


Figure 9.5: Step 4

From the Markov calculations we can obtain both expected holding cost, and expected number of orders per unit time, together with the expected number of backorders per unit time.

Problems

9.1 Consider the following situation:

- Constant failure rate (i.e., the total demand rate of spare parts) = $\lambda = 0.01$
- Number of spare parts = s = decision variable
- The spare parts are stored in a stock and retrieved from there upon a demand
- Failed components are repaired in the workshop
- Repair rate for each components repaired = $\mu = 0.1$
- We have endless number of repair men, i.e., a repairman can always start repairing a component that comes to the workshop
- $c_U = 10\,000$ = Unavailability cost per unit of time
- $c_S = 2$ = Capital cost per unit of time to keep a unit in stock

Find the optimal value of s.

9.2 Consider the situation in Problem 9.1. Solve the problem by Markov theory. Hint: Set $m = \infty$.

9.3 Consider the situation in Problem 9.1. We now also introduce $c_M = 0.25$ equal the cost per unit time per repairman available. Use Markov theory to find the optimal value of s and m.

Chapter 10

Life cycle cost and life cycle profit

10.1 Introduction

In this chapter we will give a short introduction to life cycle cost (LCC) modelling and anaylysis in connection with prioritization of renewal projects. The term LCC is defined in IEC 60300: "LCC is the cumulative cost of a product over its life cycle". The LCC concept was first introduced in the US Army and the idea was to establish the cost of development, production and use (operation and maintenance) of military equipment. In the original use the revenues was not included in the modelling. However, in order to get a complete picture we will usually also include the possible profit of a new system or product. Hence the term 'Life Cycle Profit' has been introduced. Kawacuchi and Rausand Kawauchi and Rausand (1999) suggest a process for LCC analysis comprising the following steps:

1. Problem definition
2. Cost element definition
3. System modelling
4. Data collection
5. Cost profile development
6. Evaluation
7. Reporting

In this presentation we will focus on the cost modelling aspects, i.e. mainly step 3 in the above procedure.

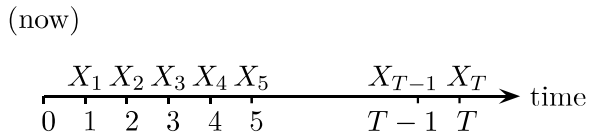


Figure 10.1: Visualisation of the cach flow, X_t

10.2 Net present value calculation

The formulas for LCC calculation are based on standard formulas used in net present value (NPV) calculations. In the following we will summarise the most frequent used formulas. The basic idea in NPV calculation is that money received in the future will be less valued than the same amount of money today. To treat this formally all future amounts are discounted to the present time, i.e. present values. We will only consider discrete time, i.e. all amounts will occur at the end of each year, or now (beginning of year one). The cash flow is illustrated in Figure 10.1.

The net present value of an amount X_t that occurs at the end of year t is:

$$\text{NPV} = X_t(1+r)^{-t} \tag{10.1}$$

where r is the discount rate. Similarly, we find the net present value of a cash flow $X_0, X_1 \ldots, X_T$:

$$\text{NPV} = \sum_{t=0}^{T} X_t(1+r)^{-t} \tag{10.2}$$

where X_0 represents in or outgoing cash now, and T is the number of years to consider.

Sometimes we want to establish the net present value of a constant yearly (nominal) amount X_A, i.e. the same amount each year. By utilising the formula for the sum of a geometric series, $\sum_{i=1}^{n} q^i = q(1-q^n)/(1-q)$ we obtain:

$$\text{NPV} = \left[\frac{1-(1+r)^{-T}}{r}\right] X_A \tag{10.3}$$

Note that NPV approaches X_A/r as T approaches infinity.

Now, consider a situation with a fixed increasing yearly value, were the first in or outgoing amount is $X_{A,v}$ (at the end of the first year), and where the amount is increasing by a factor $(1+v)$ each year. The net present value for T years is then found to be:

$$\text{NPV} = \left[\frac{1-\left(\frac{1+v}{1+r}\right)^T}{r-v}\right] X_{A,v} \tag{10.4}$$

IF $r = v$ in Equation (10.4) we use $\text{NPV} = X_{A,v}T/(1+r)$ obtained by l'Hopitals rule.

The expression in Equation (10.3) assumes that the amount $X_{A,v}$ occurs every year. In some situations we want to consider an amount $X_{A,v}$ which occurs every k year, where $k > 1$. The net present value is now given by (assuming first amount now ($t = 0$)):

$$\text{NPV} = \sum_{i=0}^{\infty} X_A (1+r)^{ki} = \frac{X_A}{1-(1+r)^{-k}} \tag{10.5}$$

If the first amount occurs at the end of year l we obtain:

$$\text{NPV} = \frac{X_A (1+r)^{-l}}{1-(1+r)^{-k}} \tag{10.6}$$

The value in Equation (10.6) assumes that we have an infinite time horizon. If we have a finite time horizon and m is the first year the yearly amount X_A vanishes, then we just subtract the contribution of those terms that vanish:

$$\text{NPV} = \frac{X_A (1+r)^{-l}}{1-(1+r)^{-k}} - \frac{X_A (1+r)^{-m}}{1-(1+r)^{-k}} \tag{10.7}$$

10.2.1 Trend modelling

When modelling trend it is important to find a simple mathematical expression for the time development. Further note that the change in the yearly amount is due to at least the following factors:

- The monetary value increases due to general conditions, such as inflation.
- The monetary value increases due to increased operating costs, e.g. physical deterioration and hence more maintenance is required.

Increased operating costs due to deterioration could usually be reset by a renewal of the system we are considering, whereas external conditions like inflation is not affected by e.g. a system renewal. In the modelling we will assume a fixed inflation rate, even if we in a more advanced model also could let the inflation rate vary. This inflation rate will be denoted v, and we could use Equation (10.4) to calculate the net present value of a amount that changes due to inflation. When we want to model increased operating cost due to deterioration, we need to introduce a local age parameter. We will let a denote the age of the system, or the age of the system since the last system renewal. When we consider degradation, we introduce the degradation rate d where we assume that the yearly increase due to deterioration equals $(1+d)$. This corresponds to an exponential growth which very often is found realistic if we have degradation mechanisms that drive the costs. Now, let c_0 be the yearly cost of operation, maintenance etc now (i.e. at time

$t = 0$). We then have the yearly cost in year t (occurring at the end of year t):

$$c_t = c_0(1+d)^t \tag{10.8}$$

In order to obtain the degradation rate d we usually need data about the costs as a function of time. A very simple approach if we know that $c(t)$ has increased by a growth factor (GF) during a period of T years. We then have:

$$d = e^{\ln(\mathrm{GF})/T} - 1 \tag{10.9}$$

10.2.2 Example areas of LCC calculations

In the following we give examples of areas where LCC analysis and calculation could be used. We differentiate between situations were decisions are related to project execution, and the progression of one project, or a portfolio of projects, and the situation where we consider which project are profitable, or how the profitability could be maximised. Examples related to project execution:

- Invest in equipment to increase efficiency in project execution, e.g., a new excavator.
- Choice between construction method A and B.
- Outsourcing of truck-maintenance.
- Lease equipment rather than own by our selves.

Examples related to project profitability:

- Development of one or more oil fields.
- Construction of a new passing loop.
- Renewal of ballast in a railway track.
- Point wise refill of ballast in order to postpone the need for a full renewal (ballast cleaning).
- Grinding of rails.
- Invest in a new production line.

There are several aspects to consider when conducting an LCC analysis, for example:

- Visualise the cost picture, enabling the possibility to work actively with eliminating the main cost drivers, or the effect of these.

- Use the LCC model as a decision support when making decision about the profitability of projects or measures, and when to conduct or implement these.
- Use the LCC model as a basis for contractual follow-up, e.g. LCC contracts.

Example 10.1
We will consider a railway system where the quality of the ballast has deteriorated during the last years, and in order to compensate for this it is proposed to do a point wise replacement of the ballast on the line. The age of the ballast is 35 years, and without this point wise refill of ballast it is expected that a full renewal (ballast cleaning) is necessary within five years. If we conduct the project we could postpone the ballast cleaning with another five year. The length of the line we are considering is 10 km. The quantities to include in the LCC model is as follows:

RC = 2.5 million Euro = Renewal cost = 250 Euro per meter for ballast cleaning.

IC = 400,000 Euro= Improvement cost, e.g. cost of point wise ballast refill.

LT = 40 years = Life length of ballast = period between ballast cleaning.

a = ballast age, i.e. effective age relative to the implemented measures. Without point wise refill of ballast $a = 35$ years, and with point wise refill of ballast $a = 30$ year. For a track that has just being renewed $a = 0$.

c_0 = 25,000 Euro = yearly cost of maintenance and operation of the track, for a new track, i.e. just being renewed.

c_{40} = 250,000 Euro = yearly cost of maintenance and operation of the track, for a track that has reached it's service life, e.g. 40 years.

$d = e^{\ln(250000/25000)/40} - 1 = 0.05925$

$c_t = c_0(1+d)^{t+a} = 25000(1+0.05925)^{t+a}$ = total maintenance and operation cost in year t (from now), and a is the effective age of the track.

r = 6% = interest rent.

We start by calculating the various LCC-terms (in million Euros) if the improvement project (point wise refill of ballast) is not executed. The total renewal cost if found by Equation (10.6):

$$\mathrm{LCC_{RC}} = \frac{\mathrm{RC}(1+r)^{-5}}{1-(1-r)^{-40}} = 2.069$$

The variable cost the next five years (up to the next renewal) is found from Equation (10.4)

$$\mathrm{LCC}_{\mathrm{VC},1} = \left[\frac{1-\left(\frac{1+d}{1+r}\right)^5}{r-d}\right] c_0(1+d)^{35} = 0.883$$

After the renewal in five year the variable costs will be reset to v_0, and then start increasing again. The net present value in one cycle is:

$$\mathrm{LCC}_{\mathrm{VC},0} = \left[\frac{1-\left(\frac{1+d}{1+r}\right)^{40}}{r-d}\right] c_0(1+d) = 0.986$$

The amount $\mathrm{LCC}_{\mathrm{VC},0}$ will then be repeated every 40 year, and the first time will be in five years:

$$\mathrm{LCC}_{\mathrm{VC},\infty} = \frac{\mathrm{LCC}_{\mathrm{VC},0}(1+r)^{-5}}{1-(1+r)^{-40}} = 0.816$$

Finally we have the total contribution from variable costs:

$$\mathrm{LCC}_{\mathrm{VC}} = \mathrm{LCC}_{\mathrm{VC},1} + \mathrm{LCC}_{\mathrm{VC},\infty} = 1.699$$

If we execute the improvement project, the calculations are similar. We start with the total renewal cost (first renewal after 10 years):

$$\mathrm{LCC}_{\mathrm{RC}} = \frac{\mathrm{RC}(1+r)^{-10}}{1-(1+r)^{-40}} = 1.546$$

The variable cost the next ten years (up to the next renewal) noting that the effective age after the improvement project is $a = 30$:

$$\mathrm{LCC}_{\mathrm{VC},1} = \left[\frac{1-\left(\frac{1+d}{1+r}\right)^{10}}{r-d}\right] c_0(1+d)^{30} = 1.322$$

After the renewal in ten year the variable costs will be reset to v_0, and then start increasing again. The net present value in one cycle, $\mathrm{LCC}_{\mathrm{VC},0}$, is the same as without the improvement project, but the first cycle will start in ten years:

$$\mathrm{LCC}_{\mathrm{VC},\infty} = \frac{\mathrm{LCC}_{\mathrm{VC},0}(1+r)^{-10}}{1-(1+r)^{-40}} = 0.610$$

Finally we have the total contribution from variable costs:

$$\mathrm{LCC}_{\mathrm{VC}} = \mathrm{LCC}_{\mathrm{VC},1} + \mathrm{LCC}_{\mathrm{VC},\infty} = 1.932$$

In this last situation we also need to include the investment cost:

$$\mathrm{LCC}_{\mathrm{IC}} = 0.4$$

Summing up all LCC contributions we find that implementing the improvement project gives a total LCC of 3.878 million versus not implementing the project gives a total cost of 3.768. Thus the improvement project is not profitable. □

Problems

10.1 Write code for implementing Equations (10.1), (10.3), (10.4) and (10.7).

10.2 Consider Example 10.1. Find the value of the discount rate r that makes the two alternatives equal from a LCC point of view. Why is this value of r higher than the initial one?

10.3 Consider Example 10.1 and investigate if ballast cleaning after 40 years is optimal. If not, find the optimal period for ballast cleaning, i.e., the optimal renewal period.

10.4 We will review the yaw motor example introduced in previous chapters. The basic model parameters are (All cost figures in kNOKs):

- $c_{\mathrm{PM}} = 15$ = cost of preventive replacement/renewal of yaw motor
- $c_{\mathrm{U}} = 66$ = total cost of failure, i.e., corrective maintenance cost and unavailability cost
- $c_{\mathrm{I}} = 2$ = inspection cost (in case of predictive maintenance)
- MTTF = 5 = mean time to failure if no maintenance is carried out
- $\alpha = 3$ = ageing parameter
- $\ell = 8$ = failure limit (Markov model, note that we use ℓ since r is used as the discount factor later on)
- $V = 3$ = increase in degradation rate (Markov model)
- $q = 0.10$ = probability that an inspection will not reveal the actual degradation

a) Find the optimal maintenance interval τ if a calendar based maintenance strategy is applied. Also find the average cost per year for this strategy.

b) Find the optimal inspection interval τ and the optimal maintenance limit m if a predictive maintenance strategy is applied (assuming the Markov model). Also find the average cost per year and compare with the calendar based strategy.

Predictive maintenance requires condition monitoring (CM) equipment to be installed. Further, the CM equipment needs upgrade every six years. This makes the predictive maintenance strategy less favourable compared to the calendar based strategy. On the other hand, the calendar based strategy will result in a larger portion of time where the yaw motor is run "close to the failure limit". Therefore bearings and other related equipment are exposed to more wear, which will result in extra cost, i.e., overhaul cost of the entire yawing system every five years. The following cost figures and parameters are to be considered in an LCC analysis

- $c_{\text{CMI}} = 50$ = Initial investment cost of condition monitoring system
- $c_{\text{CMU}} = 10$ = Upgrade cost of CM system every 6 year
- $c_{\text{O}} = 25$ = Overhaul cost of entire yawing system every 5 years
- $T = 30$ = time horizon to consider (years)
- $r = 5\%$ = discounting factor

c) Carry out an LCC analysis to determine if predictive maintenance pays off.

Chapter 11

Reliability Data Analysis

11.1 Introduction

Methods and models introduced in previous chapters require numbers for the reliability parameters such as the failure rates, repair rates, ageing parameters and so on. A parameter in this context is a quantity in a reliability model for which we assign numerical values. To obtain such numeric values several principle for parameter *estimation* exist.

11.2 Checking for trends in the data

11.2.1 Objective

In a counting process model failures are assumed to occur along the time axis, and no assumption is made regarding the status of the unit after the repair is completed. The main objective of the analysis is to reveal any trend in time, and the Nelson Aalen plot is an efficient tool.

11.2.2 Conceptual framework for counting process models

Consider one unit installed at time $t = 0$, observed over a period of time from $t = a$ to $t = b$. The recorded failure times (global or calendar time) are denoted $T_1, T_2, \ldots, T_n$. By definition $t_0 = a$. The unit is repaired after each failure, but no assumption is made about the quality of the repair. Repair times are considered neglectable. Two extremes are often considered:

- Perfect repair in which case the unit is considered "as good as new" after each repair. In this situation it is reasonable to believe in a Renewal Process (RP), and the theory of life data analysis applies. In Figure 11.1, the X_i's can be considered as the data set.

- Minimal repair in which case the unit is considered "as bad as old", i.e., the status of the component immediately before the failure occurred. In this situation it is reasonable to believe in a Non-Homogeneous Poison Process (NHPP).

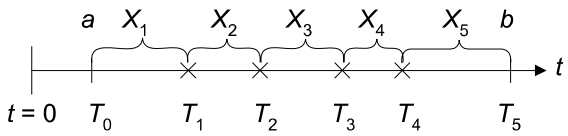


Figure 11.1: Conceptual model for a counting process

The times between each pair of failures, $X_i = T_i - T_{i-1}$ are denoted the inter-arrival times. If the inter-arrival times tend to become shorter, the system is deteriorating. On the other hand, the system is improving if the inter-arrival times tend to become longer (reliability growth). Note that any trend may be caused to both internal and external circumstances. Typical causes for improving systems are:

- Latent failures are revealed, and fixed
- Improved "organisational environment" due to gained experience of the maintenance and operational personnel
- Improved external environmental conditions
- Failed parts are replaced with new parts with higher reliability

Causes for deteriorating systems are:

- Wear-out mechanisms (of the parts)
- Aggravated external environmental conditions (e.g. more sand in the oil)
- Less resources to maintenance

11.2.3 Nelson Aalen plot

To reveal trends, the Nelson-Aalen plot is constructed. The Nelson-Aalen plot shows the cumulative number of failures on the Y-axis, and the X-axis represents the time. A convex plot indicates a deteriorating system, whereas a concave plot indicates an improving system. The idea behind the Nelson-Aalen plot is to plot the cumulative number of failures against time. We recall that the ROCOF, $w(t)$, is the failure intensity, and $W(t)$ is the expected cumulative numbers of failures in a time interval:

$$W(t) = \int_0^t (u)\,du = \mathrm{E}\,[\#\text{ of failures in } [0, t)] \tag{11.1}$$

When estimating $W(t)$ we need failure data from one or more processes (systems). Each process (system) is observed in a time interval $(a_i, b_i]$ and t_{ij} denotes failure time j in process i (global or calendar time). The information could be systematized as in Table xx

In order to construct Nelson Aalen plot the following algorithm could be used:

1. Group all the t_{ij} in Table 24, sort them, and denote the result $t_k, k = 1, 2, \ldots$..
2. For each k, let O_k denote the number of processes that are under observation just before time t_k
3. Let $\hat{W}_0 = 0$
4. Let $\hat{W}_k = \hat{W}_{k-1} + 1/O_k = 1, 2, \ldots$
5. Plot $t_k, \hat{W}_k$

Note that O_k is the number of processes that are under observation just prior to time tk, which means that the "jumps" in the estimated cumulative intensity is "adjusted" for the number of processes under observation. The points will follow a straight line if the intensity is constant. If the intensity is increasing, the t_k's will occur more and more frequent, and the cumulative plot will bend upwards (convex). If the intensity is decreasing the t_k's will occur less and less frequent, and the cumulative plot will bend downwards (concave). Figure 66 shows the Nelson-Aalen plot for the example data in Table 24.

In this presentation only the principle of maximum likelihood estimation will be addressed.

11.3 The MLE principle

The basic idea behind the Maximum Likelihood Estimation (MLE) principle is to choose the numerical values of the parameters that are the most likely ones in light of the data. The procedure goes as follows:

- Assume that we know the probability density function of the observations for which we have data. Let this distribution be denoted $f(t; \theta)$.
- The involved parameters, are unknown, and are generally denoted θ.
- We have n independent observations (data points) that we denote T_1, T_2, $\ldots$, T_n. When we refer to the actual numerical values observed, we use the (lowercase) notation t_1, t_2, $\ldots$, t_n.

The MLE principle now tells us to estimate θ by the value which is most likely given the observed data. To define "likelihood" we use the probability density function. The simultaneous probability density for $T_1, T_2, \ldots, T_n$ is given by:

$$f(t_1;\theta)f(t_2;\theta)\ldots f(t_n;\theta) = \prod_{i=1}^{n} f(t_i;\theta) \tag{11.2}$$

This density express how likely a given combination of the t-values are, given the value of θ. However, in our situation the t-values are given, whereas θ is unknown. We therefore interchange the arguments, and consider the expression as a function of θ:

$$L(\theta; t_1, t_1 \ldots, t_n) = \prod_{i=1}^{n} f(t_i;\theta) \tag{11.3}$$

where $L(\theta; t_1, t_1 \ldots t_n)$ in Equation (11.3) denotes the likelihood function. The MLE principle will now be formulated as to choose the θ-value that maximizes the likelihood function. To denote the MLE *estimator* we write a "hat" over θ, $\hat{\theta}$. Generally, θ will be a function of the observations:

$$\hat{\theta} = \hat{\theta}(T_1, T_2, \ldots, T_n) \tag{11.4}$$

When we insert numerical values for the t-values we denote the result as the parameter *estimate*.

11.3.1 Estimation in the exponential distribution

We consider the situation where we have observed n failure times, and we will estimate the failure rate, λ, under the assumption of exponentially distributed failure times.The observed failure times are denoted $t_1, t_2, \ldots, t_n$. Equation (11.3) gives:

$$L(\lambda; t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} \tag{11.5}$$

Note that the parameter is denoted λ, whereas we generally use θ. Further we denote the observations with t because we here have failure *times*. The probability density function in the exponential distribution is given by $f(t) = \lambda e^{-\lambda t}$. A common "trick" when maximising the likelihood function is to take the logarithm. Because the logarithm (ln) function is monotonically increasing, $\ln L$ will also be maximised for the same value as for which L is maximised. We could then find:

$$l(\lambda; t_1, t_2, \ldots, t_n) = \ln L(\lambda; t_1, t_2, \ldots, t_n) = n \ln \lambda - \sum_{i=1}^{n} \lambda t_i \tag{11.6}$$

By taking the derivative wrt λ and equate to zero, we easily obtain:

$$\hat{\lambda} = n/\sum_{i=1}^{n} t_i \tag{11.7}$$

11.4 How to obtain the data?

In some situations we are able to conduct experiments to get access to reliability data. We can imagine that we put n identical lightbulbs in n sockets and observe the failure times. A challenge might be that we do not have time to wait for all light bulbs to fail. This means that we will have some "real" life times and some "censored" life times. The censored life times are then the period they have survived. The fact that some light bulbs might have survived the time period of our experiment is also an information we will utilize. In the text book different types of censoring is discussed.

In most cases we do not have access to data in such a controlled manner. But very often we will have access to data from computerized maintenance management systems (CMMS) in terms of failure reports and reports from preventive maintenance.

From the CMMS it is to some extent possible to extract life time data. Several challenges are encountered in such an attempt to get life time data to use in our parameter estimation:

- Data is not reported on the appropriate level, for example we are seeking the failure rate of a pump bearing, but failures are only reported on the pump level
- There are several failure modes reported for an item, and we do not have any information regarding if the item is "as good as new" with respect to all failure modes after a corrective repair action
- Preventive maintenance is carried out, and hence we have very few "real" life times
- We have data for several items, but they are not operated under "identical" conditions, hence merging the data to get a sufficient number of data points is not easy

11.5 Failures vs censoring life times

In experiments as well as in real life there are situations where we are not able to observe the time of failure of an item. The reasons for this could be that the experiment is terminated before all items have failed, or for a real life item, the item is replaced preventively before a failure occurs. In this situations we will usually know that the item has survived a certain time period. The point of time representing this survival period is denoted a *censoring* life time. It is obvious that a censoring life time has less informative value than a real life time in order to assess the

underlying reliability parameters. However, the censoring life time represent some information we will not discard in the parameter estimation. We often put a star (*) on the censoring life times to distinguish them from the real life times. In the following we also use an indicator variable to distinguish censoring and real life times, where the value 1 means a real life time and the value 0 means a censoring life time.

11.5.1 Estimation when life times-to-failure are Weibull distributed

Now assume that we have been able to extract life time data from either controlled experiments or from our CMMS.

Let $t_1,\ t_2, \ldots t_n$ denote the observed times-to-failure including censored life times. Further let $I_1, I_2, \ldots I_n$ be indicator variables equal to one if the corresponding life time is a real life time, and equal to zero if it is a censored life time.

The censored life times are assumed to be "right censored" life times in the meaning that we know the "birth" of the item, but not the "death". The only thing we know is thus the fact that the item has survived the censored life time. To get "something" to put into the likelihood function, we then use the survivor function, $R(t)$. $R(t)$ is the likelihood that an item survives t, and this is what we need, i.e., what is "the likelihood of observing what we observed"?

Recall that the pdf of the Weibull distribution is given by $f(t;\alpha,\lambda) = \alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^{\alpha}}$ and the survivor function is given by $R(t;\alpha,\lambda) = e^{-(\lambda t)^{\alpha}}$. Thus the likelihood function is given by:

$$L(\alpha,\lambda;t_1,t_2,\ldots,I_1,I_2,\ldots) = \prod_i \left(I_i \alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^{\alpha}} + (1-I_i)e^{-(\lambda t)^{\alpha}} \right) \tag{11.8}$$

taking logarithm we obtain:

$$\begin{aligned} l(\alpha,\lambda;t_1,t_2,\ldots,I_1,I_2,\ldots) &= \ln L(\alpha,\lambda;t_1,t_2,\ldots,I_1,I_2,\ldots) \\ &= \sum_{i=1}^{n} \left(I_i \left[\ln\alpha + \alpha\ln\lambda + (\alpha-1)\ln t_i \right] - (\lambda t_i)^{\alpha} \right) \end{aligned} \tag{11.9}$$

Numerical methods are required for maximizing equation (11.9)

Example 11.1

Assume we have observed the following life times: 8,9,7,6,12,18,14,18*,6,9,11, 24,30* and 28*. Here a star (*) indicates that the life time is a censored life time. The MLE estimates are given by:

$$\hat{\alpha} \approx 1.61$$
$$\hat{\lambda} \approx 0.0555$$

obtained by the "Solver" in Excel.

11.6 Estimation in the Markov degradation model

This section presents ideas for estimation of model parameters, i.e., the degradation rates in the Markov model. Assuming the perfect state is 0 and the system runs through the states $0, 1, \ldots, r$, the objective is to find estimates for $\lambda_0, \lambda_1, \ldots, \lambda_{r-1}$ where λ_j is the transition rate from state j to state $j+1$.

Data has to be retrieved from the computerized maintenance management system. It is required that the state of the item (at the appropriate level) is reported after each inspection of the system. For system where the Markov model apply usually off-line inspections would be the information source, and these inspections take place in discrete point of times.

In the data retrieval process we need to find pair of observations/inspections where we create *one* data point. That is for each of the created data point $i, i = 1, 2, \ldots, n$ we have the triplet $< u_i, s_{1,i}, s_{2,i} >$, where u_i is the number of days between two subsequent observations, and $s_{1,i} \in \{0, 1, \ldots, r-1\}$ and $s_{2,i} \in \{0, 1, \ldots, r-1\}$ are the states for the first and second observation respectively.

11.6.1 Simple estimation procedure

A simple estimation procedure is as follows:

1. Repeat for all states $j \in \{0, 1, \ldots, r-1\}$
2. Set $f = 0$ and $t = 0$
3. Process all data points $i, i = 1, 2, \ldots, n$
   (a) If $s_{1,i} \neq s_{2,i}$ then let $f = f + 1$
   (b) Let $t = t + u_i / \left[d(s_{1,i}, s_{2,i} + 1 \right]$, where $d()$ is a distance measure between the first and second observation, i.e., the number of states between the first and second observation. For example $d(1,4) = 3$.
4. The transition rate from state j to state $j+1$ is estimated by $\hat{\lambda}_j = f/t$.

If we assume that transition times out of state j are exponentially distributed, we only need to collect the number of transitions out of state j, i.e., f and the exposure time t. Since a transition out of state j for data point i could have occurred anywhere in the interval of length u_i, we divide by the number of jumps +1, since the exposure time for that observation being in state j is $u_i / \left[d(s_{1,i}, s_{2,i}) + 1 \right]$. This last argument only holds if all transition rates are equal, but it will do for the simple estimation procedure.

11.6.2 The maximum likelihood approach

A weakness of the simple approach is that it only holds for equal transition rates. Further we do not utilize the information contain in a data point with larger jumps. The maximum likelihood approach (MLE) approach is rather simple, and it should be noted that, e.g., estimation by utilizing Markov Chain Monte Carlo simulation has been proposed and used by e.g., Bladt and Sørensen (2009) and Laskowska et al. (2023) in a similar situation.

Figure 6.1 depicts the physical states with the corresponding transition rates. The transition rates, i.e., the λ's in Figure 6.1 are above diagonal elements in the transition matrix $\mathbf{A}$, for example $a_{0,1} = \lambda_0$. The below diagonal elements are vanishing since we do not have data points representing improvements. The objective of the estimation is to obtain numerical values for $\lambda_0, \lambda_1, \ldots, \lambda_{r-1}$ as for the simple approach.

The main idea in an MLE approach is to compare the actual transitions taking place between subsequent observations with the probability of that transition should occur, given the parameters in the model, i.e, the λ-vector.

Assume that the system is in state s at time t and we consider a later point of time $t+u$ where no maintenance has been conducted in the period between. Since we know the system state at time t the $\mathbf{P}(t)$-vector is given by

$$\begin{aligned} P_s(t) &= 1 \\ P_j(t) &= 0 \text{ for } j \neq s \end{aligned} \tag{11.10}$$

For a given λ-vector in $\mathbf{A}$ we have:

$$\mathbf{P}(t+u) = \mathbf{P}(t) \cdot e^{\mathbf{A}u} \tag{11.11}$$

To calculate the exponential of a matrix, i.e., $e^{\mathbf{A}u}$ is numerically time consuming, and requires efficient library functions. The calculation in Eq. (11.11) must be carried out for each data point, and for each evaluation of the log-likelihood function, and in Section 11.6.3 below we propose an efficient approach to speed up the calculation.

To calculate Eq. (11.11) we let $t = 0$ and let $u = u_i$ for data point i. Further $s_{i,1}$ defines the starting point, i.e., when setting up $\mathbf{P}(0)$. Equation (11.12) shows the log-likelihood function:

$$l(\lambda_0, \lambda_1, \ldots) = \sum_{i=1} \ln P_{s_{i,2}}(u_i) \tag{11.12}$$

Table 11.1 shows the structure of the data needed for the ML estimation:

Table 11.1: Typical data for ML estimation

i	u_i	$s_{i,1}$	$s_{i,2}$
1	140	0	0
2	200	1	0
:			

11.6.3 Approximating matrix exponentials

In our MLE approach we need to calculate $\mathbf{P}(t) = \mathbf{P}(0)e^{\mathbf{A}t}$ for each observation each time we have to evaluate the likelihood function. Since calculating the exponential of a matrix is numerically time consuming an approximation is proposed. Recall that $\mathbf{P}(t) = \mathbf{P}(0)e^{\mathbf{A}t}$ was obtained from the Kolmogorov Markov equations:

$$\mathbf{P}(t)\cdot\mathbf{A} = \dot{\mathbf{P}}(t) \tag{11.13}$$

But rather than solving the exponential, we realize that for Δt being small, we have:

$$\mathbf{P}(t)\cdot\mathbf{A} \approx [\mathbf{P}(t+\Delta t) - \mathbf{P}(t)]/\Delta t \tag{11.14}$$

which leads to an iterative procedure where we repeatedly use:

$$\mathbf{P}(t+\Delta t) \approx \mathbf{P}(t)\,[\mathbf{A}\Delta t + \mathbf{I}] \tag{11.15}$$

and where $\mathbf{I}$ is the identity matrix. Calculating Eq. (11.15) is numerically fast. However, if we need to calculate for e.g., one year, i.e., $t = 365$ this will be time consuming. Now let $\mathbf{M}_0 = [\mathbf{A}\Delta t + \mathbf{I}]$, and calculate subsequently:

$$\mathbf{M}_i = \mathbf{M}_{i-1}\cdot\mathbf{M}_{i-1} \tag{11.16}$$

as long as $2^{i-1} < t_{\max}$. Now it follows that for $t = 1,2,4,8\ldots$, we use

$$\mathbf{P}(t) \approx \mathbf{P}(0)\cdot\mathbf{M}_i \tag{11.17}$$

where $2^{i-1} = t$. If $t \notin \{1,2,4,8\}$ let $\mathbf{b}$ be a vector for the binary representation of t, for example $\mathbf{b} = [0,1,0,1,0,0,0,...]$ corresponds to $t = 0\cdot2^0+1\cdot2^1+0\cdot2^2+1\cdot2^3+0\cdot2^4+\ldots = 0+2+0+8+0\ldots = 10$. In a similar way we calculate $\mathbf{P}(t=10) = \mathbf{P}(0)\cdot\mathbf{M}_1\cdot\mathbf{M}_3$.

Note that if the longest observation period is 2048 days, i.e., five and a half year, we only need 11 matrix multiplication to generate all the $\mathbf{M}_i$ matrices. This is done only once for each

evaluation of the likelihood function. For each observation we typically need in average five or six matrix multiplications to calculate $\mathbf{P}(t)$.

11.6.4 Including explanatory variables in the model

The elements in the λ-vector, i.e., $\lambda_{3\to 2b}, \lambda_{2b\to 2a}\ldots$ are assumed to be dependent on so-called explanatory variables also denoted stressors. If we have knowledge about the values of the explanatory variables at each defect it is possible to estimate the effect of these, i.e., the regression coefficients. If this is the case, the data in Table 11.1 is extended as indicated in Table 11.2 The

Table 11.2: Typical data for ML estimation with explanatory variables

i	u_i	$s_{i,1}$	$s_{i,2}$	$z_{i,1}$	$z_{i,2}$	...
1	140	0	0	153	86	...
2	200	0	1	137	92	...
:						

vector of explanatory variables ($\mathbf{z}_i = [z_1, z_2, \ldots]$) then represents the average value in the time interval between the two subsequent observations of each defect. Explanatory variables could be tonnage passing each year, curvature, age and type of rails and so on.

Motivated by Cox regression, see Cox (1972), we assume that each transition rate in the Markov matrix could be written on the form

$$\lambda = e^{\beta_0+\beta_1 z_1+\beta_2 z_2\ldots} \tag{11.18}$$

The impact of the regression variables may in principle vary between the various transitions, e.g., that the effect of extra load is larger as the defect grows. To simplify, we will assume that the effect are the same for all transition rates, i.e., $\lambda_{3\to 2b}, \lambda_{2b\to 2a}\ldots$. This means that for transition rate $k, k \in \{3, 2b, 2a, 1, 0\}$ we have:

$$\lambda_k = \lambda_k^0 e^{\beta_1 z_1+\beta_2 z_2\ldots} \tag{11.19}$$

where λ_k^0 is a baseline transition rate. The log-likelihood function contribution from observation i now reads:

$$l(\lambda_0^0, \lambda_1^0, \ldots, \beta_1, \beta_2, \ldots) = \sum_{i=1} \ln P_{s_{i,2}}(u_i) \tag{11.20}$$

Note that neither the λ's nor the β's are explicitly expressed in the right hand side of the log-likelihood function, but are implicitly specified first by Equation (11.19) and then by Equa-

tion (11.11)

11.7 Graphical techniques

Several graphical techniques may be used to analyse life time data, see e.g., Rausand et al. (2021). In the following we only present the total time on test (TTT) plot and the Kaplan-Meier plot. The TTT-plot gives a direct impression of the ageing parameter of time-to-failures, whereas the Kaplan-Meier plot presents the estimate for the survivor function. The Kaplan-Meier estimate can handle both real life times and censoring life times, whereas the TTT-plot can only handle complete data sets, i.e., censoring life times could not be treated.

11.8 TTT-plot

Assume we have n independent and identically distributed time-to-failure observations. The data could be obtained from items that have been operated under approximately the same conditions, and they are as good as new after a repair if we observe several failures for one item.

The observed time-to-failures are denoted $t_1, t_2, t_3, \ldots, t_n$. It may be shown that we always may sort our life time data since the ordering of collected data will in any case be arbitrary given that data are independent and identically distributed. Let $t_{(1)}, t_{(2)}, t_{(3)}, \ldots, t_{(n)}$ be the sorted time-to-failure observations, that is $t_{(1)} \leq t_{(2)} \leq t_{(3)} \leq \ldots \leq t_{(n)}$. The so-called TTT observator is defined for each point of time t as the total observed time (Total Time on Test) up to time t:

$$\mathscr{T}(t) = \sum_{j=1}^{i} t_{(j)} + (n-i)\,t \tag{11.21}$$

where i is such that $t_{(i)} \leq t < t_{(i+1)}$

The TTT plot is given by plotting the normalized TTT observator againts a normalized index, i.e.,

$$\left(\frac{i}{n}, \frac{\mathscr{T}(t_{(i)})}{\mathscr{T}(t_{(n)})} \right) \tag{11.22}$$

11.9 Kaplan-Meier plot

Let the sorted data be denoted $t_{(1)}$, $t_{(2)}$, ..., $t_{(n)}$ where also censored life times are included. Further let n_i be the number of items "under risk" at time $t_{(i)}$, i.e., the number of items still operating just prior to $t_{(i)}$. Now at time $t_{(i)}$ there might be no failure if this was a censoring time, it might be one failure, or it might even be more than one failure if two failures occurred at the

same time. Theoretically it is not possible to have two failures exactly at the same time, but due to limitation in "number of digits" to represent the failure times, we may have more than one failure at the same time. Let s_i be the number of life times observed at time $t_{(i)}$.

To obtain the Kaplan-Meier estimate we use more or less the same type of arguments as given by Rausand et al. (2021). First consider a small time interval around time $t_{(i)}$. In the beginning of this interval it will be n_i items at risk. Let p_i be the probability that one arbitrary of these items will survive this small interval. A natural estimator for p_i is given by

$$\hat{p}_i = \frac{n_{(i)} - s_{(i)}}{n_{(i)}} \tag{11.23}$$

since $n_{(i)} - s_{(i)}$ of the items we had survived this interval. Now, assume that we have an estimate, $\hat{R}_i^-$ for the probability that an item has survived *up to* the interval we are considering, then it follows that an estimate for the probability that an item will survive from $t = 0$ to *the end* of the interval is given by

$$\hat{R}_i^+ = \hat{R}_i^- \hat{p}_i \tag{11.24}$$

Following such arguments we obtain the Kaplan-Meier estimate for the survivor function at time t:

$$\hat{R}(t) = \prod_{t_{(i)} < t} \frac{n_{(i)} - s_{(i)}}{n_{(i)}} \tag{11.25}$$

Example 11.2

Assume we have observed the following life times: 8,9,7,6,12,18,14,18*,6,9,11, 24,30* and 28*. Here a star (*) indicates that the life time is a censored life time. Table 11.3 shows tje tableau for the Kaplan-Meier plot:

The following link shows the Excel file: `http://folk.ntnu.no/jvatn/eLearning/TPK4120/Excel/MLE_Kaplan_Meier.xlsx`.

11.10 Bayesian Reliability Analysis

11.10.1 Introduction

In classical estimation approaches the main idea is that we believe in "true" reliability parameters. The objective of the statistician is to "reveal" these true parameters in an "objective" manner. The statistician makes assumption regarding the observed data in terms of for example independent and identical distributed life times from some distribution class, for example the Weibull distribution. The more data available, the better will be the result.

Table 11.3: Kaplan Meier plot

t_i	I_i	n_i	s_i	$(n_i - s_i)/n_i$	$R(t_i)$
6	1	14	2	12/14=0.86	0.86
6	1	13	0	13/13=1	0.86
7	1	12	1	11/12=0.92	0.79
8	1	11	1	10/11=0.91	0.71
9	1	10	2	8/10=0.8	0.57
9	1	9	0	9/9=1	0.57
11	1	8	1	7/8=0.88	0.5
12	1	7	1	6/7=0.86	0.43
14	1	6	1	5/6=0.83	0.36
18	1	5	1	4/5=0.8	0.29
18	0	4	0	4/4=1	0.29
24	1	3	1	2/3=0.67	0.19
28	0	3	0	3/3=1	0.19
30	0	3	0	3/3=1	0.19

Bayesian methods takes another starting point. The Bayesian statistician treats the parameters as stochastic variables. Before he or she looks into the data, a subjective judgement is made about the parameters. This judgement is denoted the *prior* distribution, i.e., prior to observing the data. The prior distribution for each of the relevant parameters are described by some distribution class, for example the normal distribution, the gamma distribution and so on.

There are various ways to establish the prior distribution. The statistician may utilize statements from experts having domain knowledge relevant for the problem at hand, he or she might utilize data from similar components or systems and so forth. In this presentation we will not elaborate on how to establish the prior distribution. To find out more the key words "expert judgement" would be a starting point.

When the prior distribution is established, the statistician consider the data, $\mathbf{t}$ as *evidence*. This means that he or she will update the prior distribution to what is called the *posterior* distribution which also takes the evidence into account.

11.10.2 Procedure

The procedure for Bayesian estimation could briefly be described as follows:

1. Specify a *prior* uncertainty distribution of the reliability parameter, $\pi(\theta)$.
2. Structure reliability data information into a likelihood function, $L(\theta;\mathbf{t})$ (The likelihood function was discussed in Chapter 14 in the textbook).

3. Calculate the *posterior* uncertainty distribution of the reliability parameter vector, $\pi(\theta|\mathbf{t})$. The posterior distribution is found by $\pi(\theta|\mathbf{t}) \propto L(\theta;\mathbf{t})\pi(\theta)$, and the proportionality constant is found by requiring the posterior to integrate to one.

4. The Bayes estimate for the reliability parameter is given by the posterior mean, which in principle could be found by integration.

Note that the relation $\pi(\theta|\mathbf{t}) \propto L(\theta;\mathbf{t})\pi(\theta)$ follows from Bayes' theorem and the law of total probability: If $B_1, B_2, \ldots, B_r$ (corresponding to the θ-vector) represent a division of the sample space, and A is an arbitrary event (corresponding to $\mathbf{t}$ = the data vector), then:

$$\Pr(B_j|A) = \frac{\Pr(A|B_j)\cdot\Pr(B_j)}{\Pr(A)} = \frac{\Pr(A|B_j)\cdot\Pr(B_j)}{\sum_{i=1}^{r}\Pr(B_i)\cdot\Pr(A|B_i)}$$

Since we in the denominator sum over all possible B_i values (corresponding to the θ-vector) it will not contain θ, hence it may be regarded as a constant wrt θ. Further $\Pr(B_j)$ corresponds to the prior distribution, and $\Pr(A|B_j)$ corresponds to the likelihood function (in terms of the the product of the pdf's for each data point).

It is not obvious that we in step 4. should use the posterior *mean*. But if we aim for a single parameter estimate, and we have a posterior uncertainty distribution, it is reasonable to choose the mean value in this distribution. It might be proven that the posterior mean is the optimal value under "quadratic loss".

Example 11.3 Exponential distribution

In the following we give an example showing the main elements of the procedure. In the example we will estimate the failure rate in the constant failure rate situation. Assume that we express our prior believe[1] about the failure rate λ of a certain component (gas detector used on an oil and gas platform), in terms of the mean value $\mu = 0.7\cdot 10^{-6}$ (failures / hour), and the standard deviation $\sigma = 0.3\cdot 10^{-6}$. For mathematical convenience, it is common to choose a gamma distribution[2] with parameters α and ξ for the prior distribution. The expected value and the variance in the gamma distribution are given by $\mu = \alpha/\xi$ and $\sigma^2 = \alpha/\xi^2$ respectively, and we obtain the following expressions for α and ξ:

$$\xi = \mu/\sigma^2 = (0.7\cdot 10^{-6})/(0.3\cdot 10^{-6})^2 = 7.78\cdot 10^6$$
$$\alpha = \xi\mu \approx (7.78\cdot 10^6)\cdot(0.7\cdot 10^{-6}) \approx 5.44$$

[1] This could be based on statements from experts, see Øien et.al (1998), or by analysis of similar components (empirical Bayesian analysis).

[2] $\pi(\lambda) \propto \lambda^{\alpha-1}e^{-\xi\lambda}$ for the gamma distribution.

To establish the likelihood function, we look at the data. In this example we assume that we have observed identical units for a total time in service, t, equal to 525 600 hours (e.g., 60 detector years). In this period we have observed $n = 1$ failure. If we assume exponentially distributed time-to-failures, we know that the number of failures in a period of length $t, N(t)$, is Poisson distributed with parameter $\lambda \cdot t$. The probability of observing n failures is thus given by:

$$L(\lambda; n, t) = \Pr(N(t) = n) \propto \lambda^n \mathrm{e}^{-\lambda \cdot t}$$

and we have an expression for the likelihood function $L(\lambda; n, t)$.

The posterior distribution is found by multiplying the prior distribution with the likelihood function:

$$\pi(\lambda|n) \propto L(\lambda; n, t) \cdot \pi(\lambda) \propto \lambda^n \mathrm{e}^{-\lambda \cdot t} \cdot \lambda^{\alpha-1} \mathrm{e}^{-\xi\lambda} \propto \lambda^{(\alpha+n)-1} \mathrm{e}^{-(\xi+t)\lambda}$$

and we recognize the posterior distribution as a gamma distribution with new parameters $\alpha' = \alpha + n$, and $\xi' = \xi + t$. The Bayes estimate is given by the mean in this distribution:

$$\hat{\lambda} = \frac{\alpha + n}{\xi + t} \approx \frac{5.44 + 1}{7.78 \cdot 10^6 + 525600} \approx 0.78 \cdot 10^{-6}$$

We note that the maximum likelihood estimate gives a much higher failure rate estimate ($t/n = 1.9 \cdot 10^{-6}$), but the "weighing procedure" favours the prior distribution in our example. Generally we could interpret α and ξ here as "number of failures" and "time in service" respectively for the "prior information". Note that as more and more data becomes available, the data will dominate, and the effect of the prior distribution will be wiped out.

In Bayesian probability theory, if the posterior distribution $\pi(\theta|\mathbf{t})$ is in the same probability distribution family as the prior probability distribution $\pi(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $L(\theta; \mathbf{t})$.

A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior. If we cannot establish a conjugate prior we usually need numerical integration to solve the denominator in Bayes' theorem. The conjugate priors may also give some intuition because it shows how the data updates the prior distribution. In the example we had $\alpha' = \alpha + n$, and $\xi' = \xi + t$.

Problems

11.1 Assume we have 4 systems each with with states $0, 1, 2, \ldots, r$ where $\lambda_{i+1} = (1 + \upsilon)\lambda_{i+1}$. The systems are proof-tested every τ time units.

Numerical values are given by $r = 5, \lambda_0 = 0.001, \upsilon = 0.2$, and $\tau = 730$. The maintenance limit is $l = r - 1$. Upon a proof test (inspection to reveal state) nothing is done for states $1, 2, \ldots, l-1$. If a system is in a state $\geq l$ at a proof-test, an instantaneous repair takes place bringing the system back to state 1.

Use Monte Carlo simulation to simulate the observation set. Assume you simulate over 60 months (5 years).

11.2 In this exercise we will use the data from the previous exercise. There are only two parameters to estimate, i.e., λ_1 and υ. To get an initial guess for λ_1 we may do the following:

1. Count the number of situations in the data set where there is a transition from state 1 to another state, and let this number be denoted n_1
2. Count the number of occurrences where one system remains in state 1 from one inspection to the next inspection, or jumps from state 1 to another state from one inspection to the next inspection. Let this number be denoted t_1
3. An initial guess for λ_1 is now given by $\hat{\lambda}_1 = n_1 / t_1$

We can repeat for $\lambda_2, \lambda_3, \ldots, \lambda_{l-1}$. Note that we cannot estimate the transition rate into the fault state, i.e., λ_l by this procedure because there are no observed jumps from state l to state r due to our maintenance strategy.

We may now get an initial guess for υ by $\hat{\upsilon} = \hat{\lambda}_2 / \hat{\lambda}_1 - 1$. We could also use $\hat{\upsilon} = \hat{\lambda}_3 / \hat{\lambda}_2 - 1$ and so forth, so an average of these υ-values could be a reasonable approach to obtain an initial estimate, $\hat{\upsilon}$.

a) Calculate initial estimates for λ_1 and υ as indicated above for your simulated dataset

b) Keep υ fixed, find the LS-estimate for λ_1 according to the procedure described. Use any numerical routine for minimizing a univariate function.

c) Keep λ_1 fixed, i.e., the estimate from b), and find the LS-estimate for υ

d) Keep υ fixed, i.e., from c), and find the LS-estimate for λ_1 according to the procedure described.

e) Keep λ_1 fixed, i.e., the estimate from d), and find the LS-estimate for υ

f) Compare the result with using a minimization routine that allows for several variables.

Chapter 12

Machine learning

12.1 Introduction

Machine learning (ML) is a field of computer science that gives computers the ability to learn without being explicitly programmed. The classical textbook still used in many courses is Machine Learning by Tom M. Mitchell, Mitchell (1997). In this chapter we give a very short introduction to ML and some few examples.

☞ **Learning**: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E, Mitchell (1997).

Example 12.1 - Classification
Objects with (x,y)-coordinates are either of type A or type B. A limited training set, E, of objects is available where the correct type is known. The task, T, is to determine the type of new objects if only the (x,y) coordinates are known. As the training set increases we expect to be better in determining the type. Figure 12.2 shows an example of such data.

12.2 Type of data

The dataset essentially contain the following two types of data

- $\mathbf{x}$ = Feature vector, for example weight and height
- y = label data, for example slim or fat

A dataset of n data points is said to be *labelled* if both $\mathbf{x}$ and y are known, and is said to *unlabelled* if only $\mathbf{x}$ is known. All x-vectors is denoted the *feature set*, and all y-values is denoted the *label set* for a dataset of n data points.

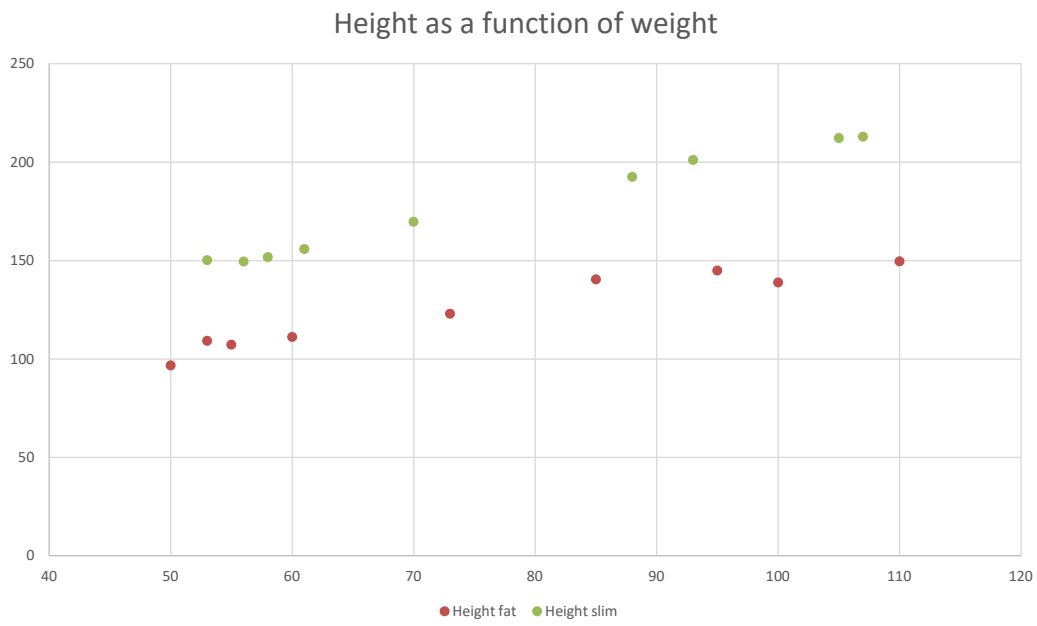


Figure 12.1: Example data - Labelling problem

In machine learning, an unknown *universal dataset* is assumed to exist, which contains all the possible data points, but this dataset is not known to us. What is available is a *training set* (training data) which is used to learn the properties and knowledge of the universal dataset

12.3 Categorization of Machine Learning

In machine learning we essentially deals with three types of learning:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

12.3.1 Supervised Learning

The training set given for supervised learning is a labelled dataset . The aim of the learing is to find the relationships between the feature set and the label set, which is the knowledge and properties we can learn from labelled dataset

- If each feature vector x corresponds to y = categorical data the learning problem is denoted as classification

- If each feature vector x corresponds to y = a real value the learning problem is defined as regression problem

12.3.2 Unsupervised Learning

The training set given for unsupervised leaning is the unlabelled dataset. The aim of could be *clustering*, probability density *estimation*, finding *association* among features, and dimensionality *reduction*. Unsupervised learning can later be input to supervised learning.

12.3.3 Reinforcement learning

Reinforcement learning is about how a computer program interacts with the environment, i.e., a real world situation like driving a car or deciding when to ask for maintenance. The computer program shall then make decisions or at least give decision support. As the time goes by the result of actions will be successful to some degree. This successfulness is then translated to a "reward" measure which the computer program tries to maximize.

12.4 Hypothesis set

Learning is about confirming hypotheses. A hypothesis, h, among other hypothesis in the hypothesis set H is a mapping function that uniquely assigns a feature vector to a label value. The objective of supervised learning is to find the best h within the set H.

It is assumed that there exist an ideal mapping function, f, which we most likely never find, but the best one found, say g, is used for future predictions

12.5 Learning algorithm

A learning algorithm, A, is required to obtain the best hypothesis within the set H. A comprises:

- An objective function, i.e., the function to be optimized for searching g
- Optimization methods used to optimize the objective function

Example 12.2 Labelling problem
Figure 12.2 shows the labelled data with three hypotheses:

- h_1 - Non-linear mapping function that assigns the label "slim" to observations above the curve, and "fat" to those below the curve

- h_2 - Linear mapping function that assigns the label "slim" to observations above the curve, and "fat" to those below the curve
- h_3 - Flat mapping function that assigns the label "slim" to observations above the curve, and "fat" to those below the curve

From the data it appears almost impossible to have a flat mapping function that uniquely separate the "slim" from the "fat", making h_3 not feasible. h_2 performs better, but having the "body mass index" (BMI) in mind, a linear classifier may not be very efficient, and therefore a non-linear mapping functions as given by hypothesis h_1 seems more efficient. To find the "best" classifier, i.e., some hypothesis g we need to define a objective function, for example based on least squares, and then fit the classifier function minimizing the square distance from the points to this function.

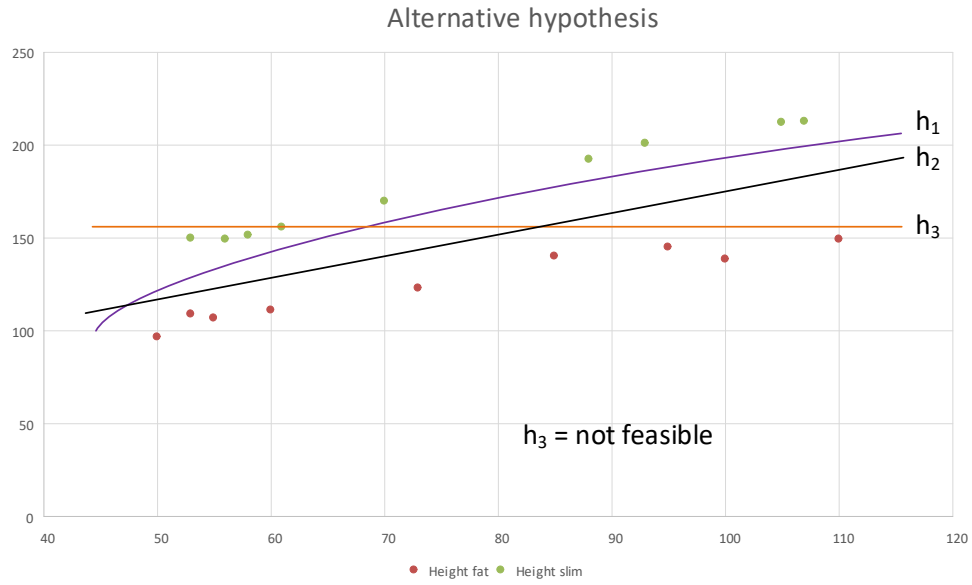


Figure 12.2: Hypothesis in the labelling problem

12.6 Support vector machines

Support vector machines (SVM) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. SVM training algorithm builds a model that assigns new observations to one category or the other, making it a non-probabilistic binary linear classifier.

Figure 12.3 shows three classifiers. The line in the middle separates "slim" and "fat" with the maximal margin. The upper line will classify new observations to "fat" if they do not follow the pattern, whereas the lower line will tend to classify them as "slim".

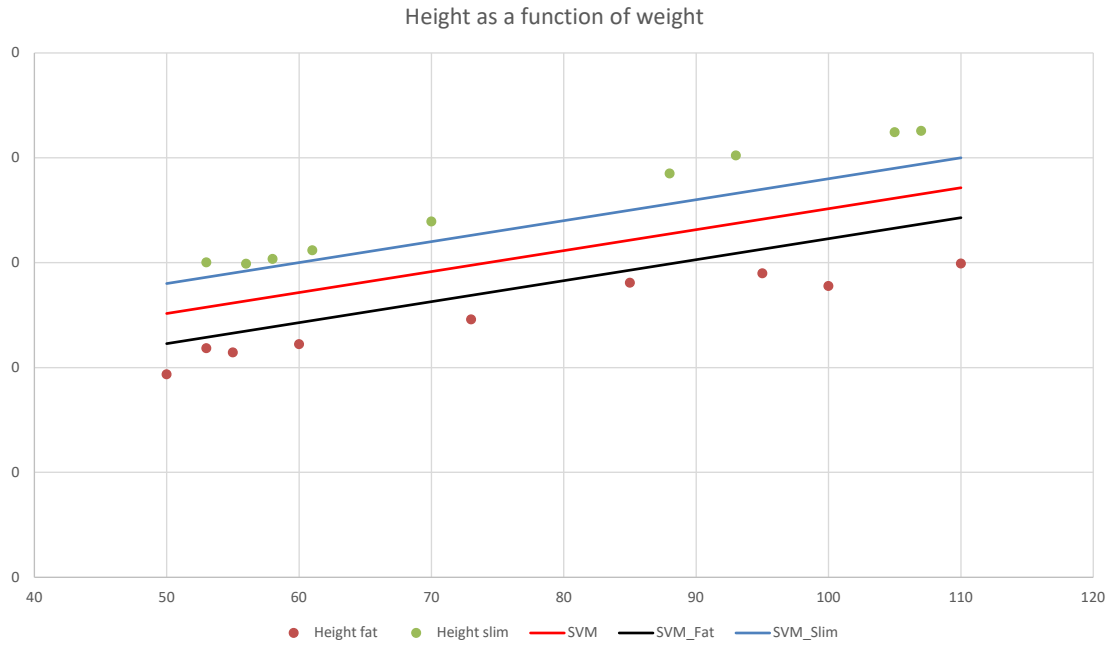


Figure 12.3: Hypothesis in the labelling problem

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the "kernel trick" which is not pursued here.

Example 12.3 Practical maintenance example

Assume that we measure the condition of a machine by the vector $\mathbf{x}$ (vibration, temperature, noise etc). Assume further that we have observed $\mathbf{x}$-values each day for one or more machines until they reach the alarm limit where we have to stop the machine.

By easy data manipulation we can then for each observation, i, calculate the number of days until the alarm/maintenance limit is reached, say d_i. In order to plan maintenance we need say D days in advance. Often D is denoted the lead time. To label the data, we now calculate y_i for each observation according to:

$$y_i = \begin{cases} \text{-1 if } d_i \leq D \\ \text{1 if } d_i > D \end{cases} \tag{12.1}$$

The feature vector $\mathbf{x}_i$ and the calculated value y_i for all observations constitute the training set. This training set is then used to obtain a classification rule. For a new observation, say $\mathbf{x}_0$ we can predict the corresponding y_0. If $y_0 = -1$ the model advocates starting maintenance.

It should be noted that if we train the model by this rather simple approach we will hopefully be quite good classifying whether whether we have sufficient time for planning maintenance if

we use the result as decision support. A weakness is obviously that we in average have a good hit, but the decision support is not very valuable since the cost of being too late is much higher than being too early. A first approach to overcome this bias, we could establish another rule for labelling the data, e.g.,

$$y_i = \begin{cases} \text{-1 if } d_i \leq D - 5 \\ \text{1 if } d_i > D - 5 \end{cases} \tag{12.2}$$

where "5" is some contingency. But it will be hard to argue for such a number. Such a classification regime is

12.7 Other learning algorithms

There are many other learning algorithms, for example

- Artificial Neural Networks (ANN)
- Deep Learning - multiple hidden layers in an artificial neural network
- Bayesian Networks
- Decision trees
- ... and many more

12.8 Artificial neural networks

Artificial neural networks (ANNs) are computing systems inspired by the biological neural networks that constitute animal brains. Compared to Support vector machines where the label is scalar, we can now have a vector of labels, i.e., the *output vector*. Figure 12.4 shows an ANN with one so-called hidden layer.

An artificial neural network is a network of simple elements called neurons. These are the nodes in the network. To the left we have a set of input nodes corresponding to the feature vector $\mathbf{x}$. To the right we have the output nodes corresponding to a vector of labelled data, i.e., a $\mathbf{y}$-vector. In the middle we have a set of hidden nodes. Figure 12.4 shows only one hidden layer, but there could be several such hidden layers which is used in so-called "deep learning" methods.

A neuron may receive input, change their internal state (= activation) according to the *input* and an *activation function* and produce some *output*

The network connects the output of certain neurons to the input of other neurons forming a directed and weighted graph:

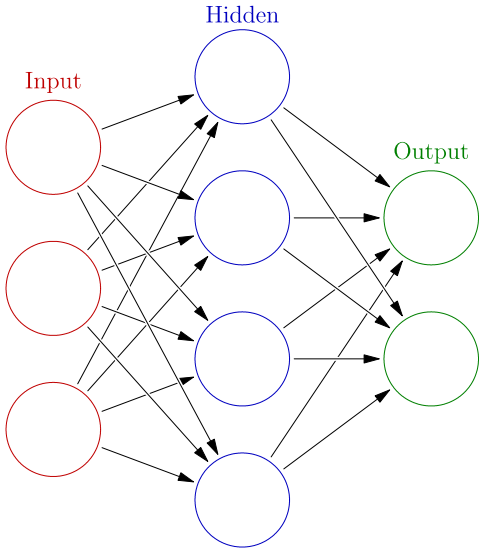


Figure 12.4: Artificial neural network (source: Wikipedia)

- Neurons = nodes item The connection between the neurons are *weighted directed edges*

Both the weights and the activation functions used contain *parameters*. Training the model means to estimate these parameters. In the estimation process an objective function is required to measure the performance of each parameter set.

12.9 Hybrid approaches

Machine learning do not require all cause-and-effects to be known (black-box), whereas probabilistic models and first principle model require an explicit model specification (grey-box and white-box). This means that ML in many cases are attractive because we do not need to work out realistic models explicitly. If we have sufficient data and we have a huge number models to be established, we can imagine that automated processes to come up with reasonable ML in short time would be possible.

However, machine learning is greedy with respect to data. Probabilistic and first principle models requires less data , and hybrid approaches could be an alternative:

- We may combine machine learning with e.g., probabilistic models
- Machine learning to search for good “health indicators”
- Use probabilistic models to assign RUL predictions (Remaining Useful Lifetime)

Example 12.4 Silicon furnace
The silicon furnaces have for many years been equipped with sensors that collect information

about various factors such as temperatures, water flow, pressure and so on. The information from these measurements are used for process control and optimization but has so far only to a small extent been used to inform maintenance decisions of the furnace equipment.

An example of furnace equipment that are monitored by sensors today are the water-cooled flexible power cables which go out to the electrode, i.e., the flexibles. The flexibles are considered as a good candidate to test the potential of using available sensor data as input to maintenance. This is because this is simple equipment with a limited number of failure modes

Assume that we on an hourly basis measure the following variables:

- p = pressure
- v = flow rate
- T = Temperature

The variables are expected to be related to each other by various physical laws, e.g., Bernoulli's equation

$$v^2/2 + gz + p/\rho = \text{constant} \tag{12.3}$$

Under normal operation the relation between the variables are expected to be rather consistent. After some time it is expected that some *chipping* takes place. The initial damage could occur according to a shock process having the PF-model in mind.

We will use simple linear regression as our ML method where we simplify and assume:

$$v = \beta_0 + \beta_1 T + \beta_2 p + \text{ error term} \tag{12.4}$$

MS Excel is a natural choice for estimating β_0, β_1 and β_2. We can easily calculate the residuals and the standard deviation of the residuals.

The anomaly detection rule is as follows:

- 3 subsequent predictions outside +/- one standard deviation

After a potential failure, we can estimate drift every day, i.e., the slope of the *deviation* between the predicted and the observed value.

12.1 The datafile for this exercise is found at folk.ntnu.no/jvatn/eLearning/MLdata/. The tasks are as follows

a) Discuss first principle models for linking the three variables, and set up a regime for early warning, i.e., anomaly detection.

b) Repeat the approach with machine learning approaches

c) Discuss what type of data you will need to train the models

Use the data under "NormalOperation" in the Excel data file to set up your model

12.2 Given that an initial damage, i.e., chipping has started, assume that the deviation between the predicted pressure and the actual pressure follows a Wiener process with constant drift.

Further, assume that chipping has been observed for several flexible power cables and "failure times" for these have been observed.

a) Discuss how you can use the result from the first principle model also for prognostics.

b) Propose a regime for estimating the RUL in this situation, given a known and fixed failure threshold

c) Use the "RunToFailure data" in the Excel file to (i) estimate the point of time when there is an anomaly, then (ii) estimate the model parameters in the Wiener process, as well as the failure threshold, assuming that there is a failure for the last observation point

d) Assume that the parameters in c) are the correct one, use the observation in RunToFailure to make RUL predictions with 90% uncertainty bands for each time steps. Compare with the "true" value.

12.10 The LS principle

The least squares (LS) principle for estimation is used when we have observations that do not come from the same distribution, but we know the expectation of each variable as a function of a set of parameters θ, and a set of explanatory variables. Previously we denoted the observations by the letter 'X', but we will now change the notation to let 'Y' denote the observations, whereas we reserve the letter 'X' for explanatory variables. We now let $\phi_i(\theta)$ denote the expectation of Y_i (the i'th observation), where the functions ϕ_i are all known, but the parameter vector θ is unknown and shall be estimated. The LS principle now states that we may estimate θ by the value that minimises the square sum of the deviations between the observed and expected values, i.e.:

$$Q(\theta) = \sum_{i=1}^{n} [y_i - \phi_i(\theta]^2 \tag{12.5}$$

Equation (12.5) is the starting point for estimating the parameters in so-called regression models. The most simple formula is given by:

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i \tag{12.6}$$

In this model x is denoted the *independent* variable, whereas Y is denoted the *dependent* variable because it depends on the independent variable, x.

The model in equation (12.6) could be extended to cover more independent variables. These are denoted regression variables , or explanatory variables. To extend the model we introduce an extra index for each x. We write x_{ij}, where index i denotes the i'th data point, whereas index j denotes the j'th explanatory variable. The model then reads:

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_r x_{i,r} \tag{12.7}$$

To obtain the LS estimators in this situation, we introduce matrix notation. Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]^{\mathrm{T}}$ be a column vector containing the dependent variables, and let $\mathbf{X}$ be the *design matrix* given by:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & ..... & x_{1r} \\ 1 & x_{21} & & x_{1r} \\ : & & x_{ij} & \\ 1 & x_{i1} & ..... & x_{nr} \end{bmatrix} \tag{12.8}$$

It could be shown that the LS estimator for $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \cdots, \beta_r]^{\mathrm{T}}$ is given as the solution of the following matrix equation:

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} \tag{12.9}$$

If the design matrix has full rank, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ will be non-singular, and the solution is given by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{12.10}$$

If one has access to a tool for matrix calculus, we easily obtain the LS estimates. We could also use commercial available statistical programs, or the "analysis" module of MS Excel.

Example 12.5 Estimation of the effects of regression variables

This example is not explicit relevant for maintenance, but given in another course. Will be updated later....

We will consider a situation where we have observed the duration of construction the foundation wall of houses. The different values are shown in the Y-column below. The variable x_1 denotes the base in square meters, whereas x_2 is an indicator of ground frost. A value is given as 1 if there is ground frost, 0 otherwise. We have also introduced the variable x_3 that denotes the walking distance from the workmen's hut to the building site:

Y	x_1	x_2	x_3
8.4	100	1	100
7.8	150	1	50
11.4	250	1	50
6.1	80	0	75
6.1	100	0	200
8.3	90	1	30
7.5	180	0	25
7.2	200	0	50
6.0	110	0	75

From MS Excel we obtain the following parameters: $\hat{\beta}_0 = 4.211$, $\hat{\beta}_1 = 0.0167$, $\hat{\beta}_2 = 2.196$, and $\hat{\beta}_3 = 0.0011$

Note that we in equation (12.7) have written the *expected* value of Y_i. Generally we write:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_r x_{i,r} + \varepsilon_i \tag{12.11}$$

where ε_i is an error-term. Very often we assume ε_i to be normally distributed, but we might also assume that ε_i is PERT distributed. To estimate the parameters in an underlying PERT distribution we calculate the predicted values:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_r x_{i,r} \tag{12.12}$$

then we estimate the error-terms by the residuals:

$$\hat{\varepsilon_i} = y_i - \hat{y}_i \tag{12.13}$$

Chapter 13

Reliabilty centred maintenance

13.1 Introduction

Reliability centred maintenance (RCM) is a method for maintenance planning developed within the aircraft industry and later adapted to several other industries and military branches. The first comprehensive documentation of RCM was given by Nowlan and Heap (1978).A major advantage of the RCM methodology is a structured, and traceable approach to determine type of preventive maintenance. This is achieved through an explicit consideration of failure modes and failure causes. A major challenge in an RCM analysis is to limit the scope of the analysis so that it is possible to carry out the analysis within the limits of time and budget. Most implementations of RCM put main focus on the identification of maintenance tasks, but do not carry out explicit optimisation of maintenance intervals. We will, however, present an approach to RCM that also enables optimisation of maintenance intervals. In order to do so, we need to structure the analysis much more than what is common in most RCM approaches.

Structuring take place at several steps in the RCM analysis. Because the failure mode and effect analysis (FMECA) is very time consuming, and because the basis for maintenance optimisation also is established through the FMECA we will introduce several means to simplify and structure this part of the analysis:

- Introduction of so-called TOP-events in the analysis. Such a TOP event could be “derailment”, “fire”, “collision train-train” for safety, and “Slow speed -40 km/h” and “Full stop” etc for punctuality. For these identified TOP events a general assessment is carried out where the total risk or cost for each such TOP event is “calculated”. The “consequence” analysis is thus reduced to totally 10-15 items, which is a very low number compared to the number of “rows” in the FMECA, which could be thousands or more.

- Introduction of generic RCM templates. A generic RCM template is the result of a general analysis of an equipment such as a turnout (mechanical part), a switch motor (electrical

part), the traction system of a train etc. In such a generic analysis we make an "average" assessment of important reliability parameters. Experience has shown that the number of "generic" RCM templates is in the order of 50, where each generic template comprise 5 to 10 "components".

- When the maintenance program is established for a specific line, or a specific train set, the generic RCM template is taken as a starting point. For this general template we make local adjustment in terms of adjustment factors. When the local adjustment factors have been defined, it is straight forward to "update" the generic template to a local analysis, where the optimisation of maintenance intervals also could be automated.
- When we know that we have several hundred or thousand physical components to treat when the maintenance program is defined, we can imagine the value of such a "generic" and "local adjustment" approach.

The RCM analysis may be carried out as a sequence of activities. Some of these activities, or steps, are overlapping in time. The RCM process comprises the following steps:

1. Study preparation
2. System selection and definition
3. Functional failure analysis (FFA)
4. Critical item selection
5. Data collection and analysis
6. Failure modes, effects and criticality analysis (FMECA)
7. Selection of maintenance actions
8. Determination of maintenance intervals
9. Preventive maintenance comparison analysis
10. Treatment of non-critical items
11. Implementation
12. In-service data collection and updating
13. Local adjustments

The various steps are discussed in the following sections with a focus on Steps 1-8. Note that the basis for step 1-12 would be the "generic approach". That is, we typically carry out these steps for "generic" systems or components, and then in step 13 we make explicit assessments reflecting the conditions related to each physical unit.

Step 1: Study preparation

The main objectives of an RCM analysis are:

1. to identify effective maintenance tasks,
2. to evaluate these tasks by some cost-benefit analysis, and
3. to prepare a plan for carrying out the identified maintenance tasks at optimal intervals.

If a maintenance program already exists, the result of an RCM analysis will often be to eliminate inefficient maintenance tasks.

Before an actual RCM analysis is initiated, an RCM project group should be established. The RCM project group should include at least one person from the maintenance function and one from the operations function, in addition to an RCM specialist.

In Step 1 "Study preparation" the RCM project group should define and clarify the objectives and the scope of the analysis. Requirements, policies, and acceptance criteria with respect to safety and environmental protection should be made visible as boundary conditions for the RCM analysis.

The part of the plant to be analysed is selected in Step 2. The type of consequences to be considered should, however, be discussed and settled on a general basis in Step 1. Possible consequences to be evaluated may comprise:

- (i) risk to humans,
- (ii) environmental damages,
- (iii) delays and cancellation of travels,
- (iv) material losses or equipment damage,
- (v) loss of marked shares, etc.

The possible consequence classes can not be measured in one common unit. It is therefore necessary to prioritise between means affecting the various consequence classes. Such a prioritisation is not an easy task and will not be discussed in this presentation. The trade-off problems can to some extent be solved within a decision theoretical framework (Vatn 1995 and Vatn *et al.* 1996).

RCM analyses have traditionally concentrated on PM strategies. It is, however, possible to extend the scope of the analysis to cover topics like corrective maintenance strategies, spare part inventories, logistic support problems, etc. The RCM project group must decide what should be part of the scope and what should be outside. The resources that are available for the analysis are usually limited. The RCM group should therefore be sober with respect to what to look into, realizing that analysis cost should not dominate potential benefits.

In many RCM applications the plant already has effective maintenance programs. The RCM project will therefore be an upgrade project to identify and select the most effective PM tasks, to recommend new tasks or revisions, and to eliminate ineffective tasks. Then apply those changes within the existing programs in a way that will allow the most efficient allocation of resources.

When applying RCM to an existing PM program, it is best to utilise, to the greatest extent possible, established plant administrative and control procedures in order to maintain the structure and format of the current program. This approach provides at least three additional benefits:

(i) It preserves the effectiveness and successfulness of the current program.

(ii) It facilitates acceptance and implementation of the project"s recommendations when they are processed.

(iii) It allows incorporation of improvements as soon as they are discovered, without the necessity of waiting for major changes to the PM program or analysis of every system.

Since we are heading for a sound basis for interval optimisation, we will need an explicit quantification of the risk associated with each "TOP event". On a general basis, we therefore need to establish the relevant risk models, both with respect to safety and punctuality. See Chapter 11 for a preliminary assessment of these risks. It is not the maintenance department that is responsible for establishing these "generic" risk models. Usually risk analyses, or safety cases exists, and these could be used as a basis for the appropriate structuring of the risk picture.

Step 2: System selection and definition

Before a decision to perform an RCM analysis is taken, two questions should be considered:

- To which systems are an RCM analysis beneficial compared with more traditional maintenance planning?
- At what level of assembly (plant, system, subsystem ...) should the analysis be conducted?

Regarding the first question, all systems may in principle benefit from an RCM analysis. With limited resources, we must, however, usually make priorities, at least when introducing the RCM approach for the first time. We should start with the systems that we assume will benefit most from the analysis. The following criteria may be used to prioritise systems for an RCM analysis:

(i) The failure effects of potential system failures must be significant in terms of safety, environmental consequences, production loss, or maintenance costs.

(ii) The system complexity must be above average.

(iii) Reliability data or operating experience from the actual system, or similar systems, should be available.

Most operating plants have developed an assembly hierarchy, i.e., an organization of the system hardware elements into a structure that looks like the root system of a tree. In the offshore oil and gas industry this hierarchy is usually referred to as the tag number system. In railway infrastructure maintenance it is common to use the disciplinary areas as the highest level in the plant register, typically we have:

- Superstructure
- Substructure
- Signalling
- Telecommunications
- Power supply (overhead line with supporting systems)
- Low voltage systems

For the rolling stock we similarly have a system breakdown:

- The breaking system including automatic train protection (ATP)
- The traction system
- The door system with interlocking connections to traction system
- The pantograph with supporting system
- The bogie system
- The coupler system
- The wagon
- The locomotive

The following terms will be used for the levels of the assembly hierarchy:

Plant: A logical grouping of systems that function together to provide an output or product by processing and manipulating various input raw materials and feed stock. An offshore gas production platform may e.g. be considered as a plant. For railway application a plant might be a maintenance area, where the main function of that "plant" is to ensure satisfactiory infrastructure functionality in that area.

System: A logical grouping of subsystems that will perform a series of key functions, which often can be summarized as one main function, that are required of a plant (e.g. feed water, steam supply, and water injection). The compression system on an offshore gas production platform may e.g. be considered as a system. Note that the compression system may consist of several compressors with a high degree of redundancy. Redundant units performing the same main function should be included in the same system. It is usually easy to identify the systems in a plant, since they are used as logical building blocks in the design process.

The system level is usually recommended as the starting point for the RCM process. This means that on an offshore oil/gas platform the starting point of the analysis should be for example the compression system, the water injection system or the fire water system, and not the whole platform. In railway application the systems were defined above as the highest level in the plant hierarchy.

The systems may be further broken down in subsystems, and subsubsystems, etc. For the purpose of the RCM-process the lowest level of the hierarchy should be what we will call an RCM analysis item:

RCM analysis item: A grouping or collection of components which together form some identifiable package that will perform at least one significant function as a stand-alone item (e.g. pumps, valves, and electric motors). For brevity, an RCM analysis item will in the following be called an analysis item. By this definition a shutdown valve, for example, is classified as an analysis item, while the valve actuator is not. The actuator is a supporting equipment to the shutdown valve, and only has a function as a part of the valve. The importance of distinguishing the analysis items from their supporting equipment is clearly seen in the FMECA in Step 6. If an analysis item is found to have no significant failure modes, then none of the failure modes or causes of the supporting equipment are important, and therefore do not need to be addressed. Similarly if an analysis item has only one significant failure mode then the supporting equipment only needs to be analyzed to determine if there are failure causes that can affect that particular failure mode. Therefore only the failure modes and effects of the analysis items need to be analysed in the FMECA in Step 6. An analysis item is usually repairable, meaning that

it can be repaired without replacing the whole item. In the offshore reliability database (OREDA) the analysis item is called an equipment unit. The various analysis items of a system may be at different levels of assembly. On an offshore platform, for example, a huge pump may be defined as an analysis item in the same way as a small gas detector. If we have redundant items, e.g. two parallel pumps, each of them should be classified as analysis items. When we in Step 6 of the RCM process identify causes of analysis item failures, we will often find it suitable to attribute these failure causes to failures of items on an even lower level of indenture. The lowest level is normally referred to as components.

Component: The lowest level at which equipment can be disassembled without damage or destruction to the items involved. Some authors refers to this lowest level as Least Replaceable Assembly (LRA), while OREDA uses the term maintainable item. It is very important that the analysis items are selected and defined in a clear and unambiguous way in this initial phase of the RCM-process, since the following analysis will be based on these analysis items. If the OREDA database is to be used in later phases of the RCM process, it is recommended as far as possible to define the analysis items in compliance with the "equipment units" in OREDA.

Step 3: Functional failure analysis (FFA)

The objectives of this step are:

(i) to identify and describe the systems required functions,

(ii) to describe input interfaces required for the system to operate, and

(iii) to identify the ways in which the system might fail to function.

Step 3(i): Identification of system functions

The objective of this step is to identify and describe all the required functions of the system. In is often recommended that the various functions are expressed in the same way, as a statement comprising a verb plus a noun - for example, "close flow", "contain fluid", "transmit signal".

A complex system will usually have a high number of different functions. It is often difficult to identify all these functions without a checklist. The checklist or classification scheme of the various functions presented below may help the analyst in identifying the functions. The same scheme will be used in Step 6 to identify functions of analysis items. The term item is therefore used in the classification scheme to denote either a system or an analysis item.

- *Essential functions:* These are the functions required to fulfil the intended purpose of the item. The essential functions are simply the reasons for installing the item. Often an es-

sential function is reflected in the name of the item. An essential function of a pump is for example to pump a fluid.

- *Auxiliary functions:* These are the functions that are required to support the essential functions. The auxiliary functions are usually less obvious than the essential functions, but may in many cases be as important as the essential functions. Failure of an auxiliary function may in many cases be more critical than a failure of an essential function. An auxiliary function of a pump is for example containment of the fluid.

- *Protective functions:* The functions intended to protect people, equipment and the environment from damage and injury. The protective functions may be classified according to what they protect, as:

- safety functions

- environment functions

- hygiene functions

An example of a protective function is the protection provided by a rupture disk on a pressure vessel, e.g., a separator.

1. *Information functions:* These functions comprise condition monitoring, various gauges and alarms etc.

2. *Interface functions:*These functions apply to the interfaces between the item in question and other items. The interfaces may be active or passive. A passive interface is for example present when an item is a support or a base for another item.

3. *Superfluous functions:* According to Moubray (1991) "Items or components are sometimes encountered which are completely superfluous. This usually happens when equipment has been modified frequently over a period of years, or when new equipment has been over specified". Superfluous functions are sometimes present when the item has been designed for an operational context that is different from the actual operational context. In some cases failures of a superfluous function may cause failure of other functions.

For analysis purposes the various functions of an item may also be classified as:

(a) *On-line functions:*These are functions operated either continuously or so often that the user has current knowledge about their state. The termination of an on-line function is called an evident failure.

(b) *Off-line functions:*These are functions that are used intermittently or so infrequently that their availability is not known by the user without some special check or test. The protective functions are very often off-line functions. An example of an off-line function is the essential function of an emergency shutdown (ESD) system on an oil platform. Many of the protective functions are off-line functions. The termination of an off-line function is called a hidden failure.

Note that this classification of functions should only be used as a checklist to ensure that all relevant functions are revealed. Discussions about whether a function should be classified as "essential" or "auxiliary" etc. should be avoided. Also note that the classification of functions here is used at the system level. Later the same classification of functions is used in the failure modes, effects and criticality analysis (FMECA) in Step 6 at the analysis item level. The system may in general have several operational modes (e.g., running, and standby), and several functions for each operating state.

The essential functions are often obvious and easy to establish, while the other functions may be rather difficult to reveal.

Step 3(ii): Functional block diagrams

The various system functions identified in Step 3(i) may be represented by functional diagrams of various types. A popular approach is the structured analysis and design technique (SADT). The SADT uses the concept of a function block with five main elements:

- Function: Definition of the function to be performed
- Input: The energy, materials, and information that are necessary to perform the function
- Control: The controls and other elements that constrain or govern how the function will be carried out
- Mechanism: The people,system, facilities or equipment necessary to carry out the function
- Output: The result of the function

Figure 13.1 illustrates this in relation to the traction system of a bike:

The necessary inputs to a function are illustrated in the functional block diagram together with the necessary control signals and the various environmental stressors that may influence the function.

It is generally not required to establish functional block diagrams for all the system functions. The diagrams are, however, often considered as efficient tools to illustrate the input interfaces

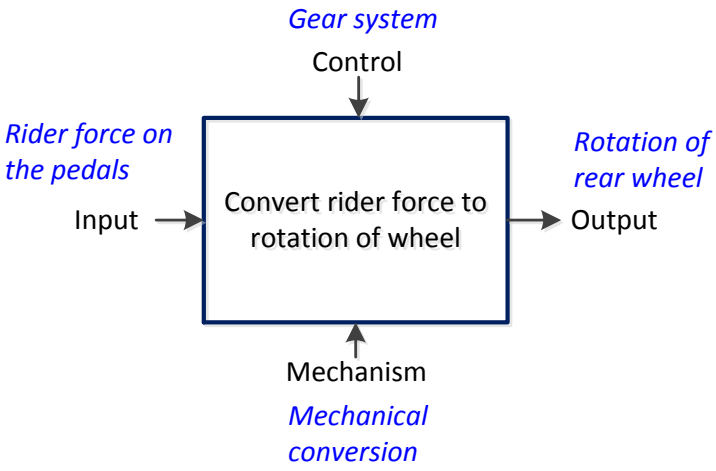


Figure 13.1: SADT for the bike example

to a function. In some cases we may want to split system functions into subfunctions on an increasing level of detail, down to functions of analysis items. The functional block diagrams may be used to establish this functional hierarchy in a pictorial manner, illustrating series-parallel relationships, possible feedbacks, and functional interfaces.

Step 3(iii): System failure modes

The next step of the FFA is to identify and describe how the various system functions may fail. In Section 4.1.5 a failure mode is defined as "The manner in which a failure occurs, independent of the cause of the failure". It is important to realize that a failure mode is a manifestation of the failure as seen from the outside, i.e. the termination of one or more functions.

In most of the RCM references the system failure modes are denoted "functional failures". Failure modes / functional failures may be classified in three main groups related to the function of the item (component, system etc):

- *Total loss of function*: In this case a function is not achieved at all, or the quality of the function is far beyond what is considered as acceptable.
- *Partial loss of function*: This group may be very wide, and may range from the nuisance category almost to the total loss of function.
- *Erroneous function*: This means that the item performs an action that was not intended, often the opposite of the intended function.

The system failure modes (functional failures) may be recorded on a specially designed FFA-form, that is rather similar to a standard FMECA form. Figure 13.2 shows an example of an FFA form.

System: Traction system Performed by: Jørn Vatn
Ref. drawing no.:#123 Date: 2023-10-18 Page: 1 of: 9

Operational mode	Function	Function requirements	Functional failure	Frequency	Criticality			
					S	E	A	C
Running	Convert rider force to rotation of wheel	Possibility to have different exchange, i.e., use of gears	No torque on back wheel	LL	-	-	H	M
			Cannot change gear	L	-	-	M	L

Figure 13.2: Part of the FFA form for the bike example

In the first column of Figure 13.2 the various operational modes of the system are recorded. For each operational mode, all the relevant functions of the system are recorded in column 2. The performance requirements to the functions, like target values and acceptable deviations are listed in column 3. For each system function (in column 2) all the relevant system failure modes are listed in column 4. In column 5 a criticality ranking of each system failure mode (functional failure) in that particular operational mode is given. The reason for including the criticality ranking is to be able to limit the extent of the further analysis by disregarding insignificant system failure modes. For complex systems such a screening is often very important in order not to waste time and money.

The criticality ranking depends on both the frequency/probability of the occurrence of the system failure mode, and the severity of the failure. The severity must be judged at the plant level.

The severity ranking should be given in the four consequence classes; (S) safety of personnel, (E) environmental impact, (A) production availability, and (C) economic losses. For each of these consequence classes the severity should be ranked as for example (H) high, (M) medium, or (L) low. How we should define the borderlines between these classes, will depend on the specific application.

If at least one of the four entries are (M) medium or (H) high, the severity of the system failure mode should be classified as significant, and the system failure mode should be subject to further analysis.

The frequency of the system failure mode may also be classified in the same three classes. (H) high may for example be defined as more than once per 5 years, and (L) low less than once per 50 years. As above the specific borderlines will depend on the application. The frequency classes may be used to prioritise between the significant system failure modes.

If all the four severity entries of a system failure mode are (L) low, and the frequency is also (L) low, the criticality is classified as insignificant, and the system failure mode is disregarded in the further analysis. If, however, the frequency is (M) medium or (H) high the system failure mode should be included in the further analysis even if all the severity ranks are (L) low, but with a lower priority than the significant system failure modes.

Step 4: Critical item selection

The objective of this step is to identify the analysis items that are potentially critical with respect to the system failure modes (functional failures) identified in Step 3(iii). These analysis items are denoted functional significant items (FSI). Note that some of the less critical system failure modes have been disregarded at this stage of the analysis. Further, the two failure modes "total loss of function" and "partial loss of function" will often be affected by the same items (FSIs).

For simple systems the FSIs may be identified without any formal analysis. In many cases it is obvious which analysis items that have influence on the system functions. For complex systems with an ample degree of redundancy or with buffers, we may need a formal approach to identify the functional significant items. If failure rates and other necessary input data are available for the various analysis items, it is usually a straightforward task to calculate the relative importance of the various analysis items based on a fault tree model or a reliability block diagram. A number of importance metrics are discussed in Appendix **??**.

The main reason for performing this task is to screen out items that are more or less irrelevant for the main system functions, i.e., in order not to waste time and money analysing irrelevant items.

In addition to the FSIs, we should also identify items with high failure rate, high repair costs, low maintainability, long lead time for spare parts, or items requiring external maintenance personnel. These analysis items are denoted maintenance cost significant items (MCSI). The sum of the functional significant items and the maintenance cost significant items are denoted maintenance significant items (MSI).

In some cases it may be beneficial to focus on critical items, in other cases we should analyse all items.

In the RCM project for the Norwegian Railway Administration the use of generic RCM analyses made it possible to analyse all identified MSIs. Thus this step tend to be less critical if a generic approach is taken.

In the FMECA analysis of Step 6, each of the MSIs will be analysed to identify their possible impact upon failure on the four consequence classes: (S) safety of personnel, (E) environmental impact, (A) production availability (punctuality), and (C) economic losses. This analysis is partly inductive and will focus on both local and system level effects.

Step 5: Data collection and analysis

The purpose of this step is to establish a basis for both the qualitative analysis (relevant failure modes and failure causes), and the quantitative analysis (reliability parameters such as MTTF, PF intervals and so on).

Step 6: Failure modes, effects and criticality analysis

The objective of this step is to identify the dominant failure modes of the MSIs identified during Step 4. The FMECA methodology is discussed in Appendix C.

Step 7: Selection of Maintenance Actions

This phase is the most novel compared to other maintenance planning techniques. A decision logic is used to guide the analyst through a question-and-answer process. The input to the RCM decision logic is the dominant failure modes from the FMECA in Step 6. The main idea is for each dominant failure mode to decide whether a preventive maintenance task is suitable, or it will be best to let the item deliberately run to failure and afterwards carry out a corrective maintenance task. There are generally three reasons for doing a preventive maintenance task:

1. to prevent a failure
2. to detect the onset of a failure
3. to discover a hidden failure

Only the dominant failure modes are subjected to preventive maintenance. To obtain appropriate maintenance tasks, the failure causes or failure mechanisms should be considered. The idea of performing a maintenance task is to prevent a failure mechanism to cause a failure. Hence, the failure mechanisms behind each of the dominant failure modes should be entered into the RCM decision logic to decide which of the following basic maintenance tasks that is applicable:

1. Continuous on-condition task (CCT)
2. Scheduled on-condition task (SCT)
3. Scheduled overhaul (SOH)
4. Scheduled replacement (SRP)
5. Scheduled function test (SFT)
6. Run to failure (RTF)

Continuous on-condition task (CCT) is a continuous monitoring of an item to find any potential failures. An on-condition task is applicable only if it is possible to detect reduced failure resistance for a specific failure mode from the measurement of some quantity.

Example:

A distance gauge on the turnout might be used to measure the distance between the switch point and stock rail to detect that the 3mm limit will be reached. At a predefined level (i.e. 2.7 mm), the system alerts the maintenance crew, which carry out an appropriate maintenance action. □

Scheduled on-condition task (SCT) is a scheduled inspection of an item at regular intervals to find any potential failures. There are three criteria that must be met for an on-condition task to be applicable:

1. It must be possible to detect reduced failure resistance for a specific failure mode.
2. It must be possible to define a potential failure condition that can be detected by an explicit task.
3. There must be a reasonable consistent age interval between the time of potential failure and the time of failure.

Examples:

A manual inspection every second month will reveal whether the “3 mm limit” is soon being reached. Appropriate maintenance action can be issued. Ultrasonic inspection of rails every year to detect cracks in the rails. □

There are two disadvantage of a scheduled versus a continuous on-condition task:

- The man-hour cost of inspection is often larger than the cost of installing the sensor
- Since the scheduled inspection is carried out at fixed points of time, one might “miss” situations where the degradation is faster than anticipated.

An advantage of a scheduled on-condition task is that the human operator is then able to “sense” information that a physical sensor will not be able to detect. This means that traditional “Walk around checks” should not be totally skipped even if sensors are installed.

Scheduled overhaul (SOH) is a scheduled overhaul of an item at or before some specified age limit, and is often called “hard time maintenance”. An overhaul task can be considered applicable to an item only if the following criteria are met (Nowlan Heap 1978):

1. There must be an identifiable age at which the item shows a rapid increase in the item's failure rate function.
2. A large proportion of the units must survive to that age.
3. It must be possible to restore the original failure resistance of the item by reworking it.

Examples:

Rehabilitation of wooden sleepers borings every three year. Lubrication of the char/slideplate every three day. Cleaning every month. □

*Scheduled replacement (SRP)*is scheduled discard of an item (or one of its parts) at or before some specified age limit. A scheduled replacement task is applicable only under the following circumstances (Nowlan Heap 1978):

1. The item must be subject to a critical failure.
2. Test data must show that no failures are expected to occur below the specified life limit.
3. The item must be subject to a failure that has major economic (but not safety) consequences.
4. There must be an identifiable age at which the item shows a rapid increase in the failure rate function.
5. A large proportion of the units must survive to that age.

Example:

Replacement of the motor every one year The motor is then either overhauled to "a god as new" condition, or replaced in the maintenance depot. □

*Scheduled function test (SFT)*is a scheduled inspection of a hidden function to identify any failure. A scheduled function test task is applicable to an item under the following conditions (Nowlan Heap 1978):

1. The item must be subject to a functional failure that is not evident to the operating crew during the performance of normal duties.
2. The item must be one for which no other type of task is applicable and effective.

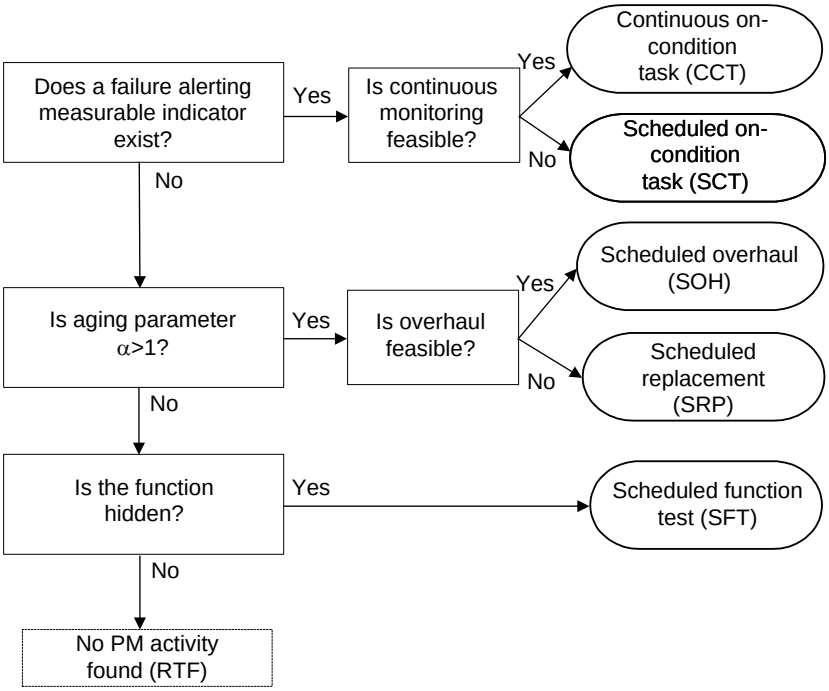


Figure 13.3: Maintenance Task Assignment/Decision logic

Example:

Sighting or hammer blow every year to detect loose lockspikes fastening chars/baseplates on wooden sleepers. □

*Run to failure (RTF)*is a deliberate decision to run to failure because the other tasks are not possible or the economics are less favourable. In many situations one maintenance task may prevent several failure mechanisms. Hence in some situations it is better to put failure modes rather than failure mechanisms into the RCM decision logic.

Note also that if a failure cause for a dominant failure mode corresponds to a supporting equipment, the supporting equipment should be defined as the "item" to be entered into the RCM decision logic.

The criteria given for using the various tasks should only be considered as guidelines for selecting an appropriate task. A task might be found appropriate even if some of the criteria are not fulfilled.

Figure 13.3 shows the RCM decision logic. Note that this logic is much simpler than those found in standard RCM references, e.g., Moubray (1991). It should be emphasized that such a logic can never cover all situations. For example in the situation of a hidden function with ageing failures, a combination of scheduled replacements and function tests is required.

1.8 Step 8: Determination of Maintenance Intervals

Usually formalised methods for optimisation of maintenance interval is not a part of the RCM analysis. In order to optimise maintenance intervals we need to structure the analysis in such a way that it fits into the maintenance optimisation models that exists. See Chapter 11 for a discussion of determination of maintenance intervals using optimisation models.

Step 9: Preventive maintenance comparison analysis

Two overriding criteria for selecting maintenance tasks are used in RCM. Each task selected must meet two requirements:

- It must be applicable
- It must be effective

Applicability: meaning that the task is applicable in relation to our reliability knowledge and in relation to the consequences of failure. If a task is found based on the preceding analysis, it should satisfy the Applicability criterion.

A PM task will be applicable if it can eliminate a failure, or at least reduce the probability of occurrence to an acceptable level - or reduce the impact of failures!

Cost-effectiveness: meaning that the task does not cost more than the failure(s) it is going to prevent. The PM task's effectiveness is a measure of how well it accomplishes that purpose and if it is worth doing. Clearly, when evaluating the effectiveness of a task, we are balancing the "cost" of "performing the maintenance with the cost of not performing it. In this context, we may refer to the cost as follows:

1. The "cost" of a PM task may include:
    - the risk of maintenance personnel error, e.g. "maintenance introduced failures"
    - the risk of increasing the effect of a failure of another component while the one is out of service
    - the use and cost of physical resources
    - the unavailability of physical resources elsewhere while in use on this task
    - production unavailability during maintenance
    - unavailability of protective functions during maintenance of these
    - "The more maintenance you do the more risk you will expose your maintenance personnel to"
2. On the other hand, the "cost" of a failure may include:

- the consequences of the failure should it occur (i.e., loss of production, possible violation of laws or regulations, reduction in plant or personnel safety, or damage to other equipment)
- the consequences of not performing the PM task even if a failure does not occur (i.e., loss of warranty)
- increased premiums for emergency repairs (such as overtime, expediting costs, or high replacement power cost).

Step 10: Treatment of non-MSIs

In Step 4 critical items (MSIs) were selected for further analysis. A remaining question is what to do with the items which are not analysed. For plants already having a maintenance program it is reasonable to continue this program for the non-MSIs. If a maintenance program is not in effect, maintenance should be carried out according to vendor specifications if they exist, else no maintenance should be performed.

Step 11: Implementation

A necessary basis for implementing the result of the RCM analysis is that the organizational and technical maintenance support functions are available. A major issue is therefore to ensure the availability of the maintenance support functions. The maintenance actions are typically grouped into maintenance packages, each package describing what to do, and when to do it.

Many accidents are related to maintenance work. When implementing a maintenance program it is therefore of vital importance to consider the risk associated with the execution of the maintenance work. Checklists could be used to identify potential risk involved with maintenance work:

- Can maintenance people be injured during the maintenance work?
- Is work permit required for execution of the maintenance work?
- Are means taken to avoid problems related to re-routing, by-passing etc.?
- Can failures be introduced during maintenance work?

Task analysis, see e.g., Kirwan and Ainsworth (1997), may be used to reveal the risk involved with each maintenance job.

Step 12: In-service data collection and updating

As mentioned earlier, the reliability data we have access to at the outset of the analysis may be scarce, or even second to none. In our opinion, one of the most significant advantages of RCM is that we systematically analyze and document the basis for our initial decisions, and, hence, can better utilize operating experience to adjust that decision as operating experience data is collected. The full benefit of RCM is therefore only achieved when operation and maintenance experience is fed back into the analysis process.

The process of updating the analysis results is also important due to the fact that nothing remain constant, best seen considering the following arguments:

- The system analysis process is not perfect and requires periodic adjustments.
- The plant itself is not a constant since design, equipment and operating procedures may change over time.
- Knowledge grows, both in terms of understanding how the plant equipment behaves and how technology can increase availability and reduce costs.

Reliability trends are often measured in terms of a non-constant ROCOF (rate of occurrence of failures). The ROCOF measures the probability of failure as a function of *calendar* time, or global time since the plant was put into operation. The ROCOF may change over time, but within one cycle the ROCOF is assumed to be constant. This means that analysis updates should be so frequent that the ROCOF is fairly constant within one period. Opposite to the ROCOF, the *failure rate fimctopm* is measuring the probability of failure as a function of *local* time, i.e., the time elapsed since last repair/replacement. However, the failure rate function can not be considered constant, if so there is no rationale for performing scheduled replacement/repair. The updating process should be concentrated on three major time perspectives:

- Short term interval adjustments
- Medium term task evaluation
- Long term revision of the initial strategy

The short term update can be considered as a revision of previous analysis results. The input to such an analysis is updated reliability figures either due to more data, or updated data because of reliability trends. This analysis should not require much resources, as the framework for the analysis is already established. Only Step 5 and Step 8 in the RCM process will be affected by short term updates.

The medium term update will also review the basis for the selection of maintenance actions in Step 7. Analysis of maintenance experience may identify significant failure *causes* not considered in the initial analysis, requiring an updated FMECA analysis in Step 6. The medium term update therefore affects Step 5 to 8.

The long term revision will consider all steps in the analysis. It is not sufficient to consider only the system being analysed, it is required to consider the entire plant with it's relations to the outside world, e.g., contractual considerations, new laws regulating environmental protection etc.

Generic and local RCM analysis

In principle, the RCM analysis should be conducted for *physical* units in an explicit operational context. This means that we for example conduct an RCM analysis for a given turnout at location X at line Y. For this turnout we identify all functions, failure modes etc. Then we propose a set of maintenance tasks, and finally chose the maintenance intervals based on the reliability performance parameters for that turnout, and the personnel and punctuality risk for that turnout. Now, there might be several hundreds of similar turnouts, but where both the reliability performance and the risk profile might vary, which again should ask for different maintenance intervals. The question is whether we need to repeat the entire RCM analysis for all the (similar) turnouts? The proposed answer to this question is to first conduct a *generic* RCM analysis, and then perform local adjustment to risk parameters. The following steps would then be required:

1. *Conduct a generic RCM analysis for selected components.* In this analysis we use generic, or average values of reliability parameters, and consequences parameters describing safety, punctuality, availability, environmental risk.

2. *Generic RCM database.* The results from the generic RCM analysis is stored in a generic RCM database, i.e., generic analyses for selected equipment types. These types could be e.g. a turnout, a main signal, traction system, break system etc. In the first place we might restrict ourselves to consider a broad class of e.g., turnouts (different manufactures). In a later phase we might want to refine our analysis to also consider qualitative different turnouts (with different failure modes).

3. *Selection of local analysis objects.* In the local analysis we work with a subset of the railway system. This could be for example one specific line, turnouts in the main track of one specific line, one specific train set, one specific train set operating on one specific line etc.

4. *Find an appropriate generic RCM template.* For a local analysis object, we now recall the corresponding generic RCM analysis from the RCM database. We first verify that the

generic RCM analysis object (template) is appropriate in terms of qualitative properties, i.e. the different functions, failure modes etc that are considered. At this point it might be necessary to add more failure modes, regard some failure modes etc. If this is the case, we add the "new" RCM object to the generic RCM database in order to make the generic RCM database more and more comprehensive.

5. *Adjust parameters.* At the local level we identify differences from the generic parameters used in the generic RCM database. For example a specific line might have very old turnouts, and hence the MTTF is shorter than the average MTTF. At this step of the procedure we have to consider all parameters that are involved in the optimisation model.

6. *Re-run the optimisation procedure.* Based on the new "local" parameters we will re-run the optimisation procedure to adjust maintenance intervals taking local differences into account. To carry out this process we need a computerised tool to streamline the work.

7. *Document the results.* The results from the local analysis is stored in a local RCM database. This is a database where only the adjustment factors are documented, for example for turnouts A, B, C and D on line Y the MTTF is 30% higher than the average. Hence the maintenance interval is also reduced accordingly.

13.2 Risk based inspection

Risk based inspection (RBI) is an approach to establish an inspection strategy for a plant. The methodology is in many aspects similar to the RCM approach. Some main differences between RCM and RBI are:

- RCM is a general method that could be applied a wide range of applications, whereas RBI is a tailor-made method which only applies typically for structural elements where the degradation could be measured, i.e. by means of inspection.

- RBI manuals usually cover a wide range of inspection methods and a discussion of the applicability of the various methods in different situations.

- The RBI method is much more integrated with the risk management system than usually is the case for RCM. This means that the safety implication of failures are more explicitly treated, and risk is often quantified on a detailed level, and compared with the overall risk acceptance criteria for the plant.

DNV-RP-G101 (2021) is a reference to RBI for mehanical equipment. Wintle et al. (2001) propose the following steps in a process diagram for plant integrity management by RBI:

1. Assess the requirements for integrity management and risk based inspection
2. Define the systems, the boundaries of systems, and the equipment requiring integrity management
3. Specify the integrity management team and responsibilities
4. Assemble plant database
5. Analyse accident scenarios, deterioration mechanisms, and assess and rank risks and uncertainties
6. Develop inspection plan within integrity management strategy
7. Achieve effective and reliable examination and results
8. Assess examination results and fitness-for-service

9a. Update plant database and risk analysis, review inspection plan and set maximum intervals to next examination

9b. Repair, modify, change operating conditions

10 Audit and review integrity management process

Appendix A

Acronyms and Greek letters

A.1 Acronyms

AI Artificial intelligence

ANN Artificial neural networks

ARP Age replacement policy

BIM Buildiing information management systems

BRP Block replacement policy

CCT Continuous on-condition task

CDF Cumulative distribution function

CM Corrective maintenance

CMMS Computerized maintenance management system

CMS Condition monitoring system

DT Digital twin

EBO Expected backorder

ETA Event tree analysis

FFA Functional failure analysis

FMECA Failure mode, effect and criticality analysis

FRACAS Failure reporting, analysis, and corrective action system

FTA Fault tree analysis

HPP Homogeneous Poison process

LCC Life cycle cost

LCP Life cycle profit

LS Least squares

LT Lead time

MDT Mean down time

ML Machine learning

MLD Mean logistic delay

MLE Maximum likelihood estimation

MOCUS Method of obtaining cut sets

MRT Mean active repair time

MSI Maintenance significant item

MTBR Mean time between renewals

MTTF Mean time to failure

NHPP Non-homogeneous Poison process

NPV Net present value

PDF Probability density function

PFD Probability of failure on demand

PM Preventive maintenance

RAMS Reliability, availability, maintainability, and safety

RBD Reliability block diagram

RCM Reliability centred maintenance

RP Renewal process

RTF Run to failure

RUL Remaining useful lifetime

SADT Structured analysis and design technique

SCADA Supervisory control and data acquisition systems

SCT Scheduled on-condition task

SD Standard deviation

SFT Scheduled function test

SOH Scheduled overhaul

SRP Scheduled replacement

SVM Support vector machines

TTT Total time on test

VBA Visual basic for applications, i.e., programming language in Excel

A.2 Greek letters

Table A.1: Greek letters and interpretation

Letter	Name	Explanation
α	alpha	Ageing parameter, and wake decay factor in offshore wind
β	beta	Used for common cause factor and regression parameter vector
λ	lambda	Failure rate in the exponential distribution, intensity parameter in the Weibull distribution, and effective failure rate as function of maintenance interval
τ	tau	Maintenance interval
ϕ	phi	Structure function
Δ	Delta	Used together with t to express a small time interval, i.e., Δt
μ	mu	Used for repair rate, and mean drift in the Wiener process
σ	sigma	Volatility in the Wiener process
Γ	Gamma	Gamma function
γ	gamma	Yaw angle in offshore wind
Θ	theta	Used for the parameter vector in estimation
Φ	Phi	Used for CDF in the standard normal distribution

Appendix B

Probability theory

B.1 Basic probability notation

In this chapter basic elements of probability theory are reviewed. Readers familiar with probability theory can skip this chapter. Readers which are very unfamiliar with this topic are advised to read an introductory textbook in probability theory.

B.1.1 Event

In order to define probability, we need to work with events. Let as an example A be the event that there is an operator error in a control room. This is written:

$$A = \{\text{operator error}\}$$

An event may occur, or not. We do not know the outcome in advance prior to the experiment or a situation in the "real life". We also use the word event to denote a set of distinct events. For example the event that we get an even number when throwing a die.

B.1.2 Probability

When events are defined, the probability that the event occurs is of interest. Probability is denoted by $\Pr(\cdot)$, i.e.,

$$\Pr(A) = \text{Probability that } A \text{ occur}$$

The numeric value of $\Pr(A)$ may be found by:

- Studying the *sample space*.

- Analysing collected data.
- Look up the value in data hand books.
- "Expert judgement" Øien et al. (1998).

The *sample space* defines all possible events. As an example let A = {It is Sunday}, B = {It is Monday}, .. , G = {It is Saturday}. The sample space is then given by $S = \{A, B, C, D, E, F, G\}$.

So-called Venn diagrams are useful when we want to analyse a subset of the sample space S. A rectangle represents the entire sample space, and closed curves such as a circle are used to represent subsets of the sample space as illustrated in Figure B.1. In the following we will

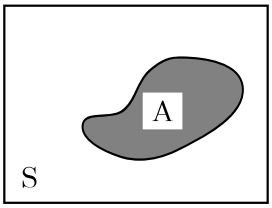
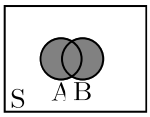
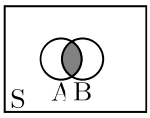


Figure B.1: Venn diagram

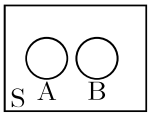
illustrate frequently used combinations of events:



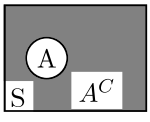
Union. We write $A \cup B$ to denote the union of A and B, i.e., the occurrence of A or B or (A and B). Let A be the event that tossing a die results in a "six", and B be the event that we get an odd number of eyes. We then have $A \cup B = \{1, 3, 5, 6\}$.



Intersection. We write $A \cap B$ to denote the intersection of A and B, i.e. the occurrence of both A and B. As an example, let A be the event that a project is not completed in due time, and let B be the event that the budget limits are exceeded. $A \cap B$ then represent the situation that the project is not completed in due time and the budget limits are exceeded.



Disjoint events. A and B are said to be *disjoint* if they can *not* occur simultaneously, i.e. $A \cap B = \emptyset$ = the empty set. Let A be the event that tossing a die results in a "six", and B be the event that we get an odd number of eyes. A and B are disjoint since they cannot occur simultaneously, and we have $A \cap B = \emptyset$.



Complementary events. The *complement* of an event A is all events in the sample space S except for A. The complement of an event is denoted by A^C. Let A be the event that tossing a die results in an odd number of eyes. A^C is then the event that we get an even number of eyes.

B.1.3 Probability and Kolmogorov's axioms

Probability is a set function Pr() which maps events A_1, $A_{2,}$... in the sample space S to real numbers. The function Pr(·) can only take values in the interval from 0 to 1, i.e. probabilities are greater or equal than 0, and less or equal than 1. Kolmogorov established the following axioms

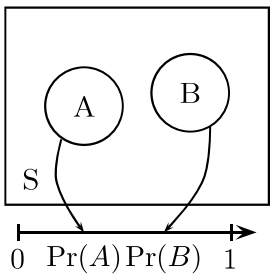


Figure B.2: Mapping of events on the interval [0,1]

which all probability rules could be derived from:

1. $0 \leq \Pr(A)$
2. $\Pr(S) = 1$
3. If A_1, A_2, A_3,... is a sequence of disjoint events we shall then have:
   $\Pr(A_1 \cup A_2 \cup \ldots) = \Pr(A_1) + \Pr(A_2) + \ldots$

The axioms are the basis for establishing calculation rules when dealing with probabilities, but they do not help us in establishing numerical values for the basic probabilities $\Pr(A_1)$, $\Pr(A_2)$, etc. Historically two lines of thoughts have been established, the classical (frequentiest) and the Bayesian approach. In the classical thinking we introduce the concept of a random experiment, where $\Pr(A_i)$ is the relative frequency with which the event A_i occurs. The probability could then be interpreted as a property of the experiment, or a property of the world. By letting nature reveal itself by doing experiments, we could in principle establish all probabilities that are of interest. Within the Bayesian framework probabilities are interpreted as subjective believe about whether A_i will occur or not. Probabilities is then not a property of the world, but rather a measure of the knowledge and understanding we have about a phenomenon.

Before we set up the basic rules for probability theory that we will need, we introduce the concepts of conditional probability and independent events.

Conditional probability. $\Pr(A|B)$ denotes the conditional probability that A will occur given that B has occurred.

Independent events. A and B are said to be *independent* if information about whether B has occurred does *not* influence the probability that A will occur, i.e., $\Pr(A|B) = \Pr(A)$.

Basic rules for probability. The following calculation rules for probability apply:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \tag{B.1}$$

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B) \text{ if A and B are independent} \tag{B.2}$$

$$\Pr(A^C) = \Pr(A \text{ does } \textit{not} \text{ occur}) = 1 - \Pr(A) \tag{B.3}$$

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{B.4}$$

Example

Let the two events A and B be defined by A = {It is Sunday} and B = {It is between 6 and 8 pm).

First we note that A and B are independent but not disjoint. We will find $\Pr(A \cap B)$, $\Pr(A \cup B)$ and $\Pr(A|B)$

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B) = \frac{1}{7} \cdot \frac{2}{24} = \frac{1}{84}$$

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = \frac{1}{7} + \frac{2}{24} - \frac{1}{84} = \frac{9}{42}$$

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/84}{2/24} = \frac{1}{7}$$

□

B.1.4 The law of total probability

In many situations it is easier to assess the probability of an event B conditionally on some other events, say $A_1, A_2, \ldots, A_r$, than unconditionally. The law of total probability could then be used to assess the unconditional probability. Now, we say that $A_1, A_2, \ldots, A_r$ is a division of the sample space if the union of all A_i's covers the entire sample space, i.e. $\mathrm{A}_1 \cup \mathrm{A}_2 \cup \ldots \cup \mathrm{A}_r = \mathrm{S}$ and the A_i's are pair wise disjoint, i.e. $\mathrm{A}_i \cap \mathrm{A}_j = \emptyset$ for $i \neq j$. An example is shown in Figure B.3.

Let $A_1, A_2, \ldots, A_r$ represent a division of the sample space S, and let B be an arbitrary event in S. The law of total probability now states:

$$\Pr(B) = \sum_{i=1}^{r} \Pr(A_i) \cdot \Pr(B|A_i) \tag{B.5}$$

Example

A special component type is ordered from two suppliers A_1 and A_2. Experience has shown that

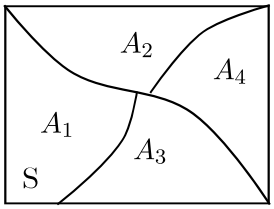


Figure B.3: Divition of the sample space

components from supplier A_1 has a defect probability of 1%, whereas components from supplier A_2 has a defect probability of 2%. In average 70% of the components are provided by supplier A_1. Assume that all components are put on a common stock, and we are not able to trace the supplier for a component in the stock. A component is now fetched from the stock, and we will calculate the defect probability, $\Pr(B)$:

$$\Pr(B) = \sum_{i=1}^{r} \Pr(A_i) \cdot \Pr(B|A_i) = \Pr(A_1) \cdot \Pr(B|A_1) + \Pr(A_2) \cdot \Pr(B|A_2) = \\ 0.7 \cdot 0.01 + 0.3 \cdot 0.02 = 1.3\%$$

□

B.1.5 Bayes theorem

Now consider the example above, and assume that we have got a defect component from the stock (event B). We will derive the probability that the component originates from supplier A_1. We then use Bayes formula that states if $A_1, A_2, \ldots, A_r$ represent a division of the sample space, and B is an arbitrary event then:

$$\Pr(A_j|B) = \frac{\Pr(B|A_j) \cdot \Pr(A_j)}{\sum_{i=1}^{r} \Pr(A_i) \cdot \Pr(B|A_i)} \tag{B.6}$$

Example
We have

$$\Pr(A_1|B) = \frac{\Pr(B|A_1) \cdot \Pr(A_1)}{\sum_{i=1}^{r} \Pr(A_i) \cdot \Pr(B|A_i)} = \frac{0.01 \cdot 0.7}{0.013} = 0.54$$

Thus, the probability of A_1 is reduced from 0.7 to 0.54 when we know that the component is defect. The reason for this is that components from supplier A_1 are the best ones, and hence when we know that the component was defect, it is less likely that it was from supplier A_1. □

B.1.6 Stochastic variables

Stochastic variables are used to describe quantities which can not be predicted exactly. Note that the term *'random quantity'* is often used to denote a stochastic variable.

$$X \text{ is stochastic} \Leftrightarrow \text{Cannot say precisely the value } X \text{ has or will take}$$

To be more precise, a stochastic variable X is a real valued function that assigns a quantitative measure to each event e_i in the sample space S, i.e., $X = X(e_i)$. Often the underlying events, e_i are of little interest. We are only interested in the stochastic variable X measured by some means.

Examples of stochastic variables are given below:

- X = Life time of a component (continuous)
- R = Repair time after a failure (continuous)
- T = Duration of a construction project (continuous)
- C = Total cost of a renewal project (continuous)
- M = Number of delayed trains next month (discrete)
- N = Number of customers arriving today (discrete)
- S = Service time for the first customer arriving today (continuous)
- W = Maintenance and operational cost next year (continuous)

Remark: We distinguish between *continuous* and *discrete* stochastic variables. Continuous stochastic variables can take any value among the real numbers, whereas discrete variables can take only a *finite* (or countable finite) number of values. □

Cumulative distribution function (CDF). A stochastic variable X is characterized by it's *cumulative distribution function*:

$$F_X(x) = \Pr(X \leq x) \tag{B.7}$$

We use subscript X to emphasise the relation to the cumulative distribution function of the quantity X. The argument (lowercase x) states which values the stochastic variable X could take, or is of our interest. From the expression we observe that $F_X(x)$ states the probability that the random quantity X is less or equal than (the numeric value of) x. A typical distribution function is shown in Figure B.4. Note that the distribution function is strictly increasing, and $0 \leq F_X(x) \leq 1$. From $F_X(x)$ we can obtain the probability that X will be within a specified interval, $[a,b)$:

$$\Pr(a < X \leq b) = F_X(b) - F_X(a) \tag{B.8}$$

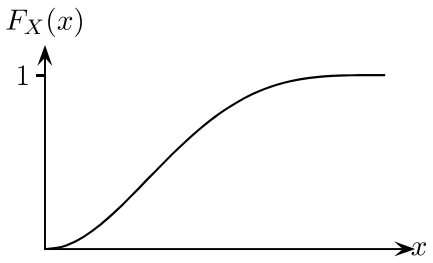


Figure B.4: Cumulative distribution function, $F_X(x)$

Note that the index X representing the stochastic variable X often is dropped if it is obvious which stochastic variable we are working with. Note also the distinction between lowercase and uppercase letters. The uppercase X is used to denote a stochastic variable, for example number of customers arriving next day. The lowercase x is just a representation of possible values X can take. For example $X = 3$.

Example

Assume that the probability distribution function of X is given by $F_X(x) = 1 - e^{-(0.01x)^2}$, and we will find the probability that X is in the interval (100,200]. From Equation (B.8) we have:

$$\Pr(100 < X \leq 200) = F_X(200) - F_X(100) =$$
$$\left[1 - e^{-(0.01 \cdot 200)^2}\right] - \left[1 - e^{-(0.01 \cdot 100)^2}\right] = e^{-1} - e^{-4} \approx 0.35$$

□

Probability density function (PDF). For a continuous stochastic variable, the *probability den-*

sity function is given by

$$f_X(x) = \frac{d}{dx} F_X(x) \tag{B.9}$$

The probability density function expresses how likely the various x-values are.

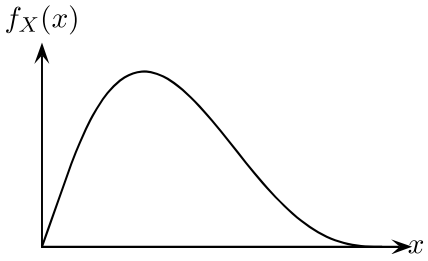


Figure B.5: Probability density function, $f_X(x)$

Note that for continuous random variables the probability that X will take a specific value vanishes. However, the probability that X will fall into a small interval around a specific value is positive. For each x-value given in Figure B.5 $f_X(x)$ could be interpreted as the probability that X will fall within a small interval around x divided by the length of this interval. Especially we have:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du \tag{B.10}$$

and

$$\Pr(a < X \leq b) = \int_{a}^{b} f_X(x)\,dx \tag{B.11}$$

The last expression is illustrated in Figure B.6.

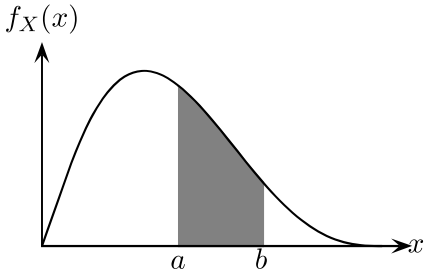


Figure B.6: The shadded area equals $\Pr(a < X \leq b)$

Random quantities that take discrete values are said to be discretely distributed. For such

quantities we introduce the point probability for X in the point x_j:

$$p(x_j) = \Pr(X = x_j) \tag{B.12}$$

where $x_1, x_2, \ldots$ are possible values X could take.

Expectation. The expectation (mean) of X is given by

$$\mathrm{E}[X] = \begin{cases} \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx & \text{if } X \text{ is continuous} \\ \sum_j x_j \cdot p(x_j) & \text{if } X \text{ is discrete} \end{cases} \tag{B.13}$$

The expectation can be interpreted as the long time run average of X, if an infinite amount of observations are available.

Median. The median of a distribution is the value m_0 of the stochastic variable X such that $\Pr(X \leq m_0) \geq 1/2$ and $\Pr(X \geq m_0) \geq 1/2$. In other words, the probability at or below m_0 is at least $1/2$, and the probability at or above m_0 is at least $1/2$.

Mode. The mode of a distribution is the value M of the stochastic variable X such that the probability density function, or point probability at M is higher or equal than for any other value of the stochastic variable. We sometimes used the term 'most likely value' rather than *mode*.

Variance. The variance of a random quantity expresses the variation in the value X will take in the long run. We denote the variance of X by:

$$\mathrm{Var}(X) = \begin{cases} \int_{-\infty}^{\infty} (x - \mathrm{E}[X])^2 \cdot f_X(x)\, dx & \text{if } X \text{ is continuous} \\ \sum_j \big((x_j - \mathrm{E}[X]\big)^2 \cdot p(x_j) & \text{if } X \text{ is discrete} \end{cases} \tag{B.14}$$

Standard deviation. The standard deviation of X is given by

$$\mathrm{SD}(X) = +\sqrt{\mathrm{Var}(X)} \tag{B.15}$$

The standard deviation defines an interval which observations are likely to fall into, i.e., if 100 observations are available, we expect that approximate[1] 67 of these observations fall in the interval $[\mathrm{E}[X] - \mathrm{SD}(X), \mathrm{E}[X] + \mathrm{SD}(X)]$.

[1] This result is valid for the normal distribution. For other distributions there may be deviation from this result.

Precision. The precision, P, is the reciprocate of the variance, i.e. $P = \frac{1}{\text{Var}(X)}$.

α-percentiles. The upper α-percentile, x_α, in a distribution $F_X(x)$ is the value satisfying $\alpha = \Pr(X > x_\alpha) = 1 - F_X(x_\alpha)$.

We end this section by giving some results regarding expectation and variances. These results apply when it is easier to express the expectation and variance of one variable if we condition on the value of another variable.

Result

Double expectation
Let X and Y be stochastic variables. We then have:

$$\text{E}[X] = \text{E}[\text{E}[X|Y]] \tag{B.16}$$

$$\text{Var}(X) = \text{E}[\text{Var}(X|Y)] + \text{Var}(\text{E}[X|Y]) \tag{B.17}$$

□

It follows easily that

$$\text{E}[X] = \text{E}[X|B]\Pr(B) + \text{E}\left[X|B^C\right]\Pr(B^C) \tag{B.18}$$

$$\begin{aligned}\text{Var}(X) &= \text{Var}(X|B)\Pr(B) + \text{Var}(X|B^C)\Pr(B^C) \\ &+ (\text{E}[X|B] - \text{E}[X])^2\Pr(B) + \left(\text{E}\left[X|B^C\right] - \text{E}[X]\right)^2\Pr(B^C)\end{aligned} \tag{B.19}$$

B.2 Common probability distributions

In this section we will present some common probability distributions. We write $X \sim$ <Name of distribution>(<parameters>) to express that X belongs to <Name of distribution>, and with parameters <parameters>. Sometimes we also use an abbreviation for the distribution, for example we write $X \sim N(3,4)$ to express that X is normally distributed with expectation 3 and variance 4.

B.2.1 The normal distribution

X is said to be normally distributed if the probability density function of X is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{B.20}$$

where μ and σ are parameters that characterise the distribution. The mean and variance are given by:

$$\begin{aligned} \mathrm{E}[X] &= \mu \\ \mathrm{Var}(X) &= \sigma^2 \end{aligned} \tag{B.21}$$

The distribution function for X could not be written on closed from. Numerical methods are required to find $F_X(x)$. It is convenient to introduce a standardised normal distribution for this purpose. We say that U is standard normally distributed if it's probability density function is given by:

$$f_U(u) = \phi(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}} \tag{B.22}$$

We then have

$$F_U(u) = \Phi(u) = \int_{-\infty}^{u} \phi(t)\mathrm{d}t = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\mathrm{d}t \tag{B.23}$$

and we observe that the distribution function of U does not contain any parameters. We therefore only need one look-up table or function representing $\Phi(u)$. A look-up table is given in Table B.1. To calculate probabilities in the non-standardised normal distribution we use the following result:

Result

If X is normally distributed with parameters μ and σ, then

$$U = \frac{X-\mu}{\sigma} \tag{B.24}$$

is standard normally distributed. □

In many situations we are interested in calculating the "truncated expectation" $\int_{-\infty}^{a} xf(x)dx$. For the normal distribution the following result may be used:

Result

Let X be normally distributed with parameters μ and σ. We then have:

Result

$$\int_{-\infty}^{a} xf(x)dx = \mu\Phi\left(\frac{a-\mu}{\sigma}\right) - \sigma\phi\left(\frac{a-\mu}{\sigma}\right) \tag{B.25}$$

where $\Phi()$ and $\phi()$ are the CDF and PDF for the standard normal distribution respectively. □

To prove Equation (B.25) first introduce $u = (x-\mu)/\sigma$ yielding $\int_{-\infty}^{a} xf(x)dx = \int_{-\infty}^{(a-\mu)/\sigma}(\sigma u - \mu)\phi(u)du$. The $\mu\phi(u)$ part of the integral is directly found by the $\Phi()$ function whereas for the $\sigma u\phi(u)$ part introduce $z = -u^2/2$ yielding $-\sigma/\sqrt{2\pi}\int_{-\infty}^{(a-\mu)^2/2\sigma^2} e^{-z}dz$. The result then follows.

Example **Calculation in the normal distribution**
Let X be normally distributed with parameters $\mu = 5$ and $\sigma = 3$. We will find $\Pr(3 < X \leq 6)$. We have:

$$\Pr(3 < X \leq 6) = \Pr(\frac{3-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{6-\mu}{\sigma}) = \Pr(\frac{3-5}{3} < U \leq \frac{6-5}{3})$$
$$= \Phi\left(\frac{1}{3}\right) - \Phi\left(\frac{-2}{3}\right) = \Phi(0.33) - (1-\Phi(0.67)) = 0.629 - 1 + 0.749 = 0.378$$

□

Problem

Consider the example in Example B.2.1, and carry out the calculations. □

Problem

Let X be the height of men in a population, and assume X is normally distributed with parameters $\mu = 181$ and $\sigma = 4$. How large percentage of the population is more than 190 cm? □

B.2.2 The exponential distribution

X is said to be exponentially distributed if the probability density function of X is given by:

$$f_X(x) = \lambda e^{-\lambda x} \tag{B.26}$$

The cumulative distribution function is given by:

$$F_X(x) = 1 - e^{-\lambda x} \tag{B.27}$$

and the mean and variance are given by:

$$\begin{aligned} \mathrm{E}[X] &= 1/\lambda \\ \mathrm{Var}(X) &= 1/\lambda^2 \end{aligned} \tag{B.28}$$

Note that for the exponential distribution, X will always be greater than 0. The parameter λ is often denoted the intensity in the distribution Example
We will obtain the probability that X is greater than it's expected value. We then have:

$$\Pr(X > \mathrm{E}[X]) = 1 - \Pr(X \leq \mathrm{E}[X]) = 1 - F_X(\mathrm{E}[X]) = e^{-\lambda \mathrm{E}[X]} = e^{-1} \approx 0.37$$

□

B.2.3 The Weibull distribution

X is said to be Weibull distributed if the probability density function of X is given by:

$$f_X(x) = \alpha\lambda(\lambda x)^{\alpha-1} e^{-(\lambda x)^\alpha} \tag{B.29}$$

The cumulative distribution function is given by:

$$F_X(x) = 1 - e^{-(\lambda x)^\alpha} \tag{B.30}$$

and the mean and variance are given by:

$$\begin{aligned} \mathrm{E}[X] &= \frac{1}{\lambda}\Gamma\left(\frac{1}{\alpha}+1\right) \\ \mathrm{Var}(X) &= \frac{1}{\lambda^2}\left[\Gamma\left(\frac{2}{\alpha}+1\right) - \Gamma^2\left(\frac{1}{\alpha}+1\right)\right] \end{aligned} \tag{B.31}$$

where $\Gamma(\cdot)$ is the gamma function. Note that in the Weibull distribution X will always be positive.

The Weibull distribution is often used as a distribution for time to failure of components, partly because it is rather simple, and flexible. For time to failure distributions we introduce the

concept of the failure rate function. The failure rate function is given by:

$$z(t) = \frac{f(t)}{1 - F(t)} \tag{B.32}$$

here we use t for running time, and we skip the index for the name of the stochastic variable when it is obvious from the context. It follows that the failure rate function for the Weibull distribution is given by:

$$z(t) = \alpha \lambda^{\alpha} t^{\alpha - 1} \tag{B.33}$$

B.2.4 The gamma distribution

X is said to be gamma distributed if the probability density function of X is given by:

$$f_X(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} (x)^{\alpha - 1} e^{-\lambda x} \tag{B.34}$$

α is denoted the intensity parameter whereas λ is denoted the intensity parameter. For integer values of α the gamma distribution is often denoted the *Erlang* distribution. The cumulative distribution function could then be found on closed form:

$$F_X(x) = 1 - \sum_{n=0}^{\alpha - 1} \frac{(\lambda x)^{\alpha}}{n!} e^{-(\lambda x)} \tag{B.35}$$

For non-integer values of α numerical methods are required to obtain the cumulative distribution function. The mean and variance are given by:

$$\begin{aligned} \mathrm{E}[X] &= \frac{\alpha}{\lambda} \\ \mathrm{Var}(X) &= \frac{\alpha}{\lambda^2} \end{aligned} \tag{B.36}$$

If we know the expectation E and the variance V in the gamma distribution, we may obtain the parameters α and λ by: $\lambda = E/V$, and $\alpha = \lambda \cdot E$. The gamma distribution is often used as a prior distribution in a Bayesian approach.

For integer values of α the gamma distribution and in particular the Erlang distribution may be seen as a distribution for a sum of exponentially distributed stochastic variables:

Result

Let $Z_1, Z_2, \ldots Z_k$ be independent and exponentially distributed with parameter λ. The variable $X = \sum_{i=1}^{k} Z_i$ is then gamma distributed with shape parameter k and scale parameter λ. □

B.2.5 The inverted gamma distribution

X is said to be inverted gamma distributed if the probability density function of X is given by:

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha+1} e^{-\lambda/x} \tag{B.37}$$

The mean and variance are given by:

$$\begin{aligned} \mathrm{E}[X] &= \lambda/(\alpha-1) \\ \mathrm{Var}(X) &= \lambda^2(\alpha-1)^{-2}(\alpha-2)^{-1} \end{aligned} \tag{B.38}$$

Note that if X is gamma distributed with parameters α and λ, then $Y = X^{-1}$ has an inverted gamma distribution with parameters α and $1/\lambda$. If we know the expectation, E and the variance, V, of an inverted gamma distribution we could obtain α and λ by $\alpha = E^2/V + 2$, and $\lambda = E \cdot (\alpha - 1)$.

B.2.6 The lognormal distribution

X is said to be lognormally distributed if the probability density function of X is given by: eqStream: Lognormal Distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\tau} \frac{1}{x} e^{-\frac{1}{2\tau^2}(\log x - \nu)^2} \tag{B.39}$$

We write $X \sim \mathrm{LN}(\nu,\tau)$. The mean and variance of X is given by

$$\begin{aligned} \mathrm{E}[X] &= e^{\nu + \frac{1}{2}\tau^2} \\ \mathrm{Var}(X) &= e^{2\nu}(e^{2\tau^2} - e^{\tau^2}) \end{aligned} \tag{B.40}$$

The following result could be utilised:

Result

If X is lognormally distributed with parameters ν and τ, then $Y = \ln X$ is normally distributed[2] with expected value ν and variance τ^2. □

[2] $\ln(\cdot)$ is the natural logarithm function

B.2.7 The binomial distribution

Before the binomial distribution is defined, binomial trials are defined. Let A be an event, and assume that the following holds:

i) n trials are performed, and in each trial we record whether A occurrs or not.

ii) The trials are stochastic *independent* of each other.

iii) For each trial $\Pr(A) = p$

When i)-iii) is satisfied, we say that we have binomial trials. Now let X be the number of times event A occurs in such a binomial trial. X is then a stochastic variable with a binomial distribution. This is written $X \sim Bin(n, p)$.

The probability function is given by

$$\Pr(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, .., n \tag{B.41}$$

The cumulative distribution function $\Pr(X \leq x)$ is given in statistical tables. For the binomial distribution, expectation and variance are given by:

$$\begin{aligned} \mathrm{E}[X] &= np \\ \mathrm{Var}(X) &= np(1-p) \end{aligned} \tag{B.42}$$

B.2.8 The Poisson distribution

The Poisson distribution is often appropriate in the situation where the stochastic variable may take the values 0,1,2,..., and where the expected number of occurrences is proportional to an exposure measure such as time or space. For the Poisson distribution we have the following point distribution:

$$p(x) = \Pr(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \tag{B.43}$$

For the poison distribution, expectation and variance are given by:

$$\begin{aligned} \mathrm{E}[X] &= \lambda \\ \mathrm{Var}(X) &= \lambda \end{aligned} \tag{B.44}$$

It can be proved that the Poisson distribution is appropriate if the following situation applies: Consider the occurrence of a certain event (e.g., a component failure) in an interval (a, b), and assume the following:

1. A could occur anywhere in (a,b), and the probability that A occurs in $(t, t+\Delta t)$ is approximately equal to $\lambda\Delta t$, and is independent of t (Δt should be small).

2. The probability that A occurs several times in $(t, t+\Delta t)$ is approximately 0 for small values of Δt.

3. Let I_1og I_2 be disjoint intervals in (a, b). The event A occurs within I_1 is independent of if the event A occurs in I_2.

When the criteria above are fulfilled we say we have a *Poisson point process* with intensity λ. The number of occurrences (X) of A in (a, b) is then Poisson distributed with parameter $\lambda(b-a)$:

$$p(x) = \Pr(X = x) = \frac{[\lambda(b-a)]^x}{x!} e^{-\lambda(b-a)} \tag{B.45}$$

Result: Times between occurrence in the Poisson process

In a Poisson point process with parameter λ the times between the occurrence of the event A are exponentially distributed with parameter λ. □

B.2.9 The inverse-Gauss distribution

The inverse-Gauss distribution is often used when we have an "under laying" deterioration process. If this deterioration process follows a Wiener process with drift η and diffusion constant δ^2, the time[3] T, until the first time the process reaches the value ω will be Inverse-Gauss distributed with parameters $\mu = \omega/\eta$, and $\lambda = \omega^2/\delta^2$.

If the failure progression $\Omega(t)$ follows a Wiener process it could be proven that $\Omega(t)$ - $\Omega(s)$ is normally distributed with expected value $\eta(t-s)$ and variance $\delta^2(t$ - s). That is η is the average growth rate in the process, whereas δ^2 is an expression for the variation of the growth around the average value.

For the inverse-Gauss distribution we have:

$$F_T(t) = \Phi\left(\sqrt{\frac{\lambda}{t}}\left(\frac{t}{\mu}-1\right)\right)) + \Phi\left(-\sqrt{\frac{\lambda}{t}}\left(\frac{t}{\mu}+1\right)\right) e^{2\lambda/\mu} \tag{B.46}$$

and

$$\mathrm{E}[T] = \mu \tag{B.47}$$

$$\mathrm{Var}(T) = \mu^3/\lambda \tag{B.48}$$

[3] We use the symbol T rather than the more general symbol X here since this modell is so explicitly linked to the time.

B.3 Distribution of sums, products and maximum values

B.3.1 Distribution of sums

If X_1, X_2,...,X_n are random variables we might obtain the expected value, the variance and the standard deviation of the sum of the x-es:

$$\mathrm{E}[X_1 + X_2 + \ldots + X_n] = \mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{E}[X_i] \tag{B.49}$$

$$\mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \tag{B.50}$$

$$\mathrm{SD}\left(\sum_{i=1}^{n} X_i\right) = \sqrt{\sum_{i=1}^{n} [\mathrm{SD}(X_i)]^2} \tag{B.51}$$

Note that Equations (B.50) and (B.51) are only valid if the x-es are stochastically independent. If there is dependency between the x-es we need to include a covariance term, e.g., if we only have two variables X_1 and X_2 we have:

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2) \tag{B.52}$$

where $\mathrm{Cov}(X_1, X_2)$ is the covariance between X_1 and X_2.

The results above help us in determine the expectation and variance of a sum of stochastic variables, but the results could not be used to establish the probability distribution of the sum. In the following we refer some results we could utilise in many situations.

Result: Sum of normally distributed stochastic variables

Let X_1, X_2,...,X_n be independent normally distributed. Let Y be the sum of the x-es, i.e. $Y = \sum_{i=1}^{n} X_i$. Y is then normally distributed with $\mathrm{E}[Y] = \sum_{i=1}^{n} \mathrm{E}[X_i]$ and $\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$. □

Result: Sum of exponentially distributed stochastic variables

Let X_1, X_2,...,X_n independent exponentially distributed with parameter λ. Let Y be the sum of the x-es, i.e. $Y = \sum_{i=1}^{n} X_i$. Y is then gamma distributed with parameters n and λ. □

Result: Sum of gamma distributed stochastic variables

Let X_1, X_2,...,X_n independent gamma distributed with parameters α and λ. Let Y be the sum of the x-es, i.e. $Y = \sum_{i=1}^{n} X_i$. Y is then gamma distributed with parameters $n\alpha$ and λ. □

Result: Central limit theorem

Let X_1, X_2,...,X_n be a sequence of identical independent distributed stochastic variables with expected value μ and standard deviation σ. As n approaches infinity, the average value of the x-es will asymptotically have a normal distribution with expected value μ and standard deviation $\sigma/\sqrt{n}$. Similarly, the sum of the x-es will asymptotically have a normal distribution with expected value $n\mu$ and standard deviation $\sigma\sqrt{n}$. □

Several generalizations for finite variance exist which do not require identical distribution but incorporate some conditions which guarantee that none of the variables exert a much larger influence than the others. Two such conditions are the Lindeberg condition and the Lyapunov condition. Now, as n approaches infinity, the sum of the x-es will asymptotically have a normal distribution with expected value $\sum_{i=1}^{n} \mathrm{E}[X_i]$ and variance $\sum_{i=1}^{n} \mathrm{Var}(X_i)$.

B.3.2 Distribution of a product

If X_1, X_2,...,X_n are *independent* stochastic variables we might obtain the expected value, the variance and the standard deviation of the product of the x-es:

$$\mathrm{E}[X_1 \cdot X_2 \cdot \ldots \cdot X_n] = \mathrm{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathrm{E}[X_i] \tag{B.53}$$

The results for the variance and standard deviation is more complicated, and we only present the results for n=2.

$$\mathrm{Var}(X_1 X_2) = \mathrm{Var}(X_1)\mathrm{Var}(X_2) + \mathrm{Var}(X_1)\left(\mathrm{E}[X_2]\right)^2 + \mathrm{Var}(X_2)\left(\mathrm{E}[X_1]\right)^2 \tag{B.54}$$

$$\mathrm{SD}(X_1 X_2) = \sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2) + \mathrm{Var}(X_1)\left(\mathrm{E}[X_2]\right)^2 + \mathrm{Var}(X_2)\left(\mathrm{E}[X_1]\right)^2} \tag{B.55}$$

Table B.1: The Cumulative Standard Normal Distribution

$$\Phi(z) = \Pr(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, du$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.500	.504	.508	.512	.516	.520	.524	.528	.532	.536
0.1	.540	.544	.548	.552	.556	.560	.564	.567	.571	.575
0.2	.579	.583	.587	.591	.595	.599	.603	.606	.610	.614
0.3	.618	.622	.626	.629	.633	.637	.641	.644	.648	.652
0.4	.655	.659	.663	.666	.670	.674	.677	.681	.684	.688
0.5	.691	.695	.698	.702	.705	.709	.712	.716	.719	.722
0.6	.726	.729	.732	.732	.739	.742	.745	.749	.752	.755
0.7	.758	.761	.764	.767	.770	.773	.776	.779	.782	.785
0.8	.788	.791	.794	.797	.800	.802	.805	.808	.811	.813
0.9	.816	.819	.821	.824	.826	.829	.831	.834	.836	.839
1.0	.841	.844	.846	.849	.851	.853	.855	.858	.860	.862
1.1	.864	.867	.869	.871	.873	.875	.877	.879	.881	.883
1.2	.885	.887	.889	.891	.893	.894	.896	.898	.900	.901
1.3	.903	.905	.907	.908	.910	.911	.913	.915	.916	.918
1.4	.919	.921	.922	.924	.925	.926	.928	.929	.931	.932
1.5	.933	.934	.936	.937	.938	.939	.941	.942	.943	.944
1.6	.945	.946	.947	.948	.949	.951	.952	.953	.954	.954
1.7	.955	.956	.957	.958	.959	.960	.961	.962	.962	.963
1.8	.964	.965	.966	.966	.967	.968	.969	.969	.970	.971
1.9	.971	.972	.973	.973	.974	.974	.975	.976	.976	.977
2.0	.977	.978	.978	.979	.979	.980	.980	.981	.981	.982
2.1	.982	.983	.983	.983	.984	.984	.985	.985	.985	.986
2.2	.986	.986	.987	.987	.987	.988	.988	.988	.989	.989
2.3	.989	.990	.990	.990	.990	.991	.991	.991	.991	.992
2.4	.992	.992	.992	.992	.993	.993	.993	.993	.993	.994
2.5	.994	.994	.994	.994	.994	.995	.995	.995	.995	.995
2.6	.995	.995	.996	.996	.996	.996	.996	.996	.996	.996
2.7	.997	.997	.997	.997	.997	.997	.997	.997	.997	.997
2.8	.997	.998	.998	.998	.998	.998	.998	.998	.998	.998
2.9	.998	.998	.998	.998	.998	.998	.999	.999	.999	.999
3.0	.999	.999	.999	.999	.999	.999	.999	.999	.999	.999

$\Phi(-z) = 1 - \Phi(z)$

Appendix C

Failure Modes, Effects, and Criticality Analysis

C.1 Introduction

Failure Mode and Effects Analysis (FMEA) was one of the first systematic techniques for failure analysis. It was developed by reliability engineers in the late 1950's to determine problems that could arise from malfunctions of military systems. A Failure Mode and Effects Analysis is often the first step in a systems reliability study. It involves reviewing as many components, assemblies and subsystems as possible to identify possible failure modes and the causes and effects of such failures. For each component, the failure modes and their resulting effects on the rest of the system are written onto a specific FMEA form. There are numerous variations of such forms. An example of an FMEA form is shown below.

A Failure Mode and Effects Analysis is mainly a qualitative analysis, which is usually carried out during the design stage of a system. The purpose is then to identify design areas where improvements are needed to meet the reliability requirements. The Failure Mode and Effect Analysis can be carried out either by starting at the component level and expanding upwards (the "bottom up" approach), or from the system level downwards (the "top down" approach). The component level to which the analysis should be conducted is often a problem to define. It is often necessary to make compromises since the workload could be tremendous even for a system of moderate size. It is, how¬ever, a general rule to expand the analysis down to a level at which failure rate estimates are available or can be obtained. Most Failure Mode and Effects Analyses are carried out according to the "bottom-up" approach. One may, however, for some particular systems save a considerable amount of effort by adopting the "top down" approach. With this approach, the analysis is carried out in two or more stages. The first stage is an analysis on the functional block diagram level. The possible failure modes and failure effects of each functional block are identified based on knowledge of the block's required function, or from

experience on similar equipment. One then proceeds to the next stage, where the components within each functional block are analysed. If a functional block has no failure modes which are critical, then no further analysis of that block needs to be performed. By this screening, it is possible to save time and effort. A weakness of this "top down" approach lies in the fact that it is not possible to ensure that all failure modes of a functional block have been identified.

An FMEA becomes a Failure Modes, Effects and Criticality Analysis (FMECA) if practicalities or priorities are assigned to the failure mode effects.

More detailed information on how to conduct a Failure Mode and Effects Analysis (and an FMECA) may be found in:

- MIL-STD 1629 "Procedures for performing a failure mode and effect analysis"
- IEC 60812 "Procedures for failure mode and effect analysis (FMEA)"
- SAE ARP 5580 "Recommended failure modes and effects analysis (FMEA) practices for non-automobile applications"
- SAE J1739 "Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) and Effects Analysis for Machinery (Machinery FMEA)"

C.2 FMECA procedure

1. FMECA prerequisites
2. System structure analysis
3. Failure analysis and preparation of FMECA worksheets
4. Team review
5. Corrective actions

Important aspects of FMECA prerequisites are:

1. Define the system to be analysed in terms of (a) System boundaries (which parts should be included and which should not), (b) Main system missions and functions (incl. functional requirements), and (c) Operational and environmental conditions to be considered
2. Collect available information that describes the system to be analysed; including drawings, specifications, schematics, component lists, interface information, functional descriptions, and so on

3. Collect information about previous and similar designs from internal and external sources; including FRACAS (Failure reporting, analysis, and corrective action system) interviews with design personnel, operations and maintenance personnel, component suppliers, and so on.

Various methods for the system structure analysis exist. An SADT analysis may be a good starting point.

A suitable FMECA worksheet for the analysis has to be decided. In many cases the client (customer) will have requirements to the worksheet format - for example to fit into his maintenance management system. A sample FMECA worksheet covering the most relevant columns is given in Figure C.1.

FMECA

System: Performed by:
Subsystem: Date:
Function: Page:

DESCRIPTION OF UNIT			DESCRIPTION OF FAILURE			EFFECT OF FAILURE			FAILURE RATE	CRITICALITY	CORRECTIVE ACTION	REMARKS
IDENTI - FICATION	OPERATIONAL MODE	FUNCTION	FAILURE MODE	FAILURE MECHANISM	HOW TO DETECT	LOCAL	SYSTEM	OPERAT . STATUS				

Figure C.1: Relevant columns in an FMECA form.

C.3 Columns in the FMECA form

Please refer to the listed references to get a comprehensive discussion of the various columns in an FMECA form. In the following we highlight some important aspects.

C.3.1 Operational mode

Example of operational modes are: idle, standby, and running. Operational modes for an air plane include, for example, taxi, take-off, climb, cruise, descent, approach, flare-out, and roll. Also note that operational mode at the system level is not the same as operational mode at the component level.

C.3.2 Failure mechanisms and failure causes

Failure mechanisms relates to physical, chemical or other processes that deteriorates the entity, and leads to a failure The term “failure cause” is often used in two different ways:

- Proximate cause, e.g., failure on a lower level in the system hierarchy such as a defect bearing in a pump

- Root cause, for example bad maintenance, inadequate design etc.

Figure C.2 illustrates the relation between function, failure mode, failure cause and failure mechanism:

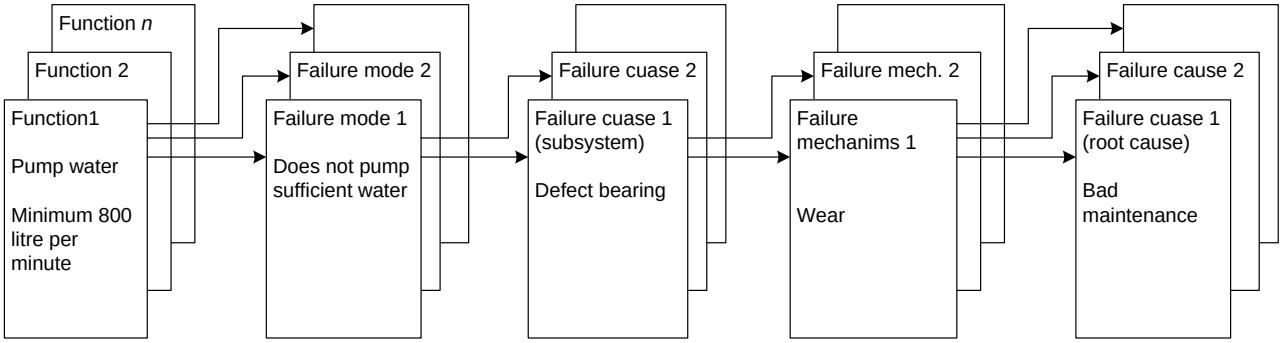


Figure C.2: Relation between function, failure mode, failure cause and failure mechanism

C.3.3 Hidden versus evident failures

We often distinguish between hidden and evident failures. The term “hidden” often relates to entities that is not continuously demanded. For example the SIFA valve on a train (bleed of the air pressure by activation) is a hidden function, and a failure will not be detected automatically. The term “evident” relates to entities that are continuously demanded, and a failure will most likely be detected immediately. Note that the same SIFA-valve will also have a evident function (“not bleed of air pressure under normal operation”) because an unintended activation immediately will be detected (breaks are activated).

C.4 Example of FMECA form

Figure C.3 shows an example FMECA form for a bike.

FMECA

System: Bike
Subsystem: Traction
Function: Convert pedal force from the rider to wheel torque

Performed by: Jørn
Date: Some date
Page: 1

DESCRIPTION OF UNIT			DESCRIPTION OF FAILURE			EFFECT OF FAILURE			FAILURE RATE	CRITICALITY	CORRECTIVE ACTION	REMARKS
IDENTI - FICATION	OPERATIONAL MODE	FUNCTION	FAILURE MODE	FAILURE MECHANISM	HOW TO DETECT	LOCAL	SYSTEM	OPERAT . STATUS				
Chain	Running	Convert torque from crank to gear	Not converting Uneven movement	Fatigue	Inspection	No gear torque	Bike is not moving	Can't reach lecture today	Low	High	Bring chain lock	

Figure C.3: Example FMECA for a bike

Appendix D

Markov Analysis

D.1 Introduction

Markov analysis is used to model systems which have many different states. These states range from "perfect function" to a total fault state. The migration between the different states may often be described by a so-called Markov-model. The possible transitions between the states may further be described by a Markov-diagram, or a state diagram.

Markov analysis is well suited for deciding reliability characteristics of a system. Especially the method is well suited for small systems with complicated maintenance strategies. In a Markov analysis the following topics will be of interest:

- The average time the system is in each state. These numbers might further form a basis for economic considerations
- How many times the system in average "visits" the various states. This information might further be used to estimate the need for spare parts, and maintenance personnel
- The mean time until the system enters one specific state, for example a critical state.

The main learning objectives of this Appendix is:

- The definition of a Markov process, and what is meant by homogeneous transition probabilities
- That it is possible to derive the Markov differential equations, but we do not need to know all the details
- The understanding of states and how we derive the Markov transition diagram
- How to map the information in the Markov transition diagram into the transition matrix, **A**

- The understanding of the Markov differential equations: $\mathbf{P}(t) \cdot \mathbf{A} = \dot{\mathbf{P}}(t)$
- How to find the steady state solution: (i) Steps required for the numerical solution, (ii) that it is possible to find an analytical solution....
- The definition of the visiting frequencies, and how to find them
- That the time dependent solution is given by $\mathbf{P}(t) = \mathbf{P}(0)e^{t\mathbf{A}}$ which cannot be used unless a comprehensive matrix library is available
- How to use the iterative scheme: $\mathbf{P}(t+\Delta t) \approx \mathbf{P}(t)[\mathbf{A}\Delta t + \mathbf{I}]$ if we have access to a computer with simple matrix functions

D.2 Definitions

A Markov process is a special type of stochastic processes where the process posses the so-called Markov property. A stochastic process $\{X(t), t \in \Theta\}$ is a collection of random variables. The set Θ is called the *index set* of the process. For each index t in Θ, $X(t)$ is called the *state* of the process at time t.

In the general presentation we always assume that $X(t)$ can only take the values $1, 2, ..., r$. In practical examples it is often convenient to allow for an additional zero value for the state variable. A process is said to have the Markov property if:

$$\Pr(X(t+s) = j | X(s) = i \cap \text{ some history up to time } s) = \Pr(X(t+s) = j | X(s) = i)$$

This means that given the process is in state i at some time s, the probability of being in another state, say j, t time units later is independent of the history up to time s, i.e., we may ignore all information regarding the process in the past when looking into the future. The only thing that counts is the current state.

This general presentation also only treats Markov processes with *stationary transition probabilities*. This means that:

$$\Pr(X(t+s) = j | X(s) = i) = \Pr(X(t) = j | X(0) = i) \text{ for all } s, t \geq 0$$

that is, the probability of going from state i to j during a time period of t is independent of the starting point of such a "journey".

The following notation is introduced:

$$P_{ij}(t) = \Pr(X(t) = j | X(0) = i)$$

The so-called sojourn time, $\tilde{T}_i$, is the time the process spends in state i from it arrives to state i before it jumps out of state i. Further let T_{ij} denote the time the process spends in state i before it eventually jumps to state j. The *transition rate* from state i to state j is denoted a_{ij} and is the limiting conditional probability of jumping to state j given that the process is in state i (divided by the length of the interval considered). It may be argued that the Markov property and the stationary transition probabilities yields that all transition times are *exponentially distributed*. The total rate of transition out of state i is denoted α_i, where

$$\alpha_i = \sum_{j \neq i} a_{ij}$$

From the fact that the sojourn time and all other transition times are exponentially distributed it follows that:

$$P_{ii}(\Delta t) = \Pr(\tilde{T}_i > \Delta t) = e^{-\alpha_i \Delta t} \approx 1 - \alpha_i \Delta t$$
$$P_{ij}(\Delta t) = \Pr(T_{ij} \leq \Delta t)) = 1 - e^{-a_{ij} \Delta t} \approx a_{ij} \Delta t$$

Rearranging and letting Δt approach 0, we get:

$$\lim_{\Delta t \to 0} \frac{1 - P_{ii}(\Delta t)}{\Delta t} = \alpha_i \tag{D.1}$$
$$\lim_{\Delta t \to 0} \frac{P_{ij}(\Delta t)}{\Delta t} = a_{ij} \tag{D.2}$$

These two equations will later be used to obtain the Kolmogorov differential equations. From the Markov property and the law of total probability we have:

$$P_{ij}(t+s) = \sum_{k=1}^{r} P_{ik}(t) P_{kj}(s)$$

This equation is denoted the Chapman-Kolmogorov equations. We utilize this equation to find:

$$P_{ij}(t + \Delta t) = P_{ij}(\Delta t + t) = \sum_{k=1}^{r} P_{ik}(\Delta t) P_{kj}(t)$$

Rearranging (having in mind we are seeking the derivative) we get:

$$P_{ij}(t + \Delta t) - P_{ij}(t) = \sum_{\substack{k=1 \\ k \neq i}}^{r} P_{ik}(\Delta t) P_{kj}(t) - [1 - P_{ii}(\Delta t)] P_{ij}(t)$$

Now dividing by Δt, inserting equations D.1 and D.2, letting $\Delta t \to 0$, and defining $a_{ii} = -\alpha_i$, we

get after some rearrangements:

$$\dot{P}_{ij}(t) = \sum_{k=1}^{r} a_{ik} P_{kj}(t) \tag{D.3}$$

These differential equations are denoted the *Kolmogorov backward equations*. Similarly, we may obtain the *Kolmogorov forward equations*:

$$\dot{P}_{ij}(t) = \sum_{k=1}^{r} a_{kj} P_{ik}(t) \tag{D.4}$$

The term 'backward' refers to that the equations were derived by considering an instant jump (transition) to state k back at the start of the interval, and then go to the required state j, i.e., first Δt and then t. The 'forward' equations are derived by first considering going from i to k during time t and then make an instant jump to the required state j at the end of the interval, i.e., first t and then Δt.

D.3 Markov state equations

We now assume that we know the initial state, and assume that the process started in state i. We then simplify notation by omitting the index for the initial state, hence we write $P_j(t)$ instead of $P_{ij}(t)$.

It is convenient to introduce matrix and vector notation. First we define the transition rate matrix, **A**:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rr} \end{bmatrix}$$

where

$$a_{ii} = -\alpha_i = -\sum_{\substack{j=1 \\ j \neq i}}^{r} a_{ij}$$

which means that the diagonal elements are defined such that the sum of each row equals zero.

Further we define the row vectors: $\mathbf{P}(t) = [P_1(t), P_2(t), \ldots, P_r(t)]$ and $\dot{\mathbf{P}}(t) = [\dot{P}_1(t), \dot{P}_2(t), \ldots, \dot{P}_r(t)]$.

We may then write the Kolmogorov forward equations on matrix format:

$$[P_1(t), P_2(t), \ldots, P_r(t)] \cdot \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rr} \end{bmatrix} = [\dot{P}_1(t), \dot{P}_2(t), \ldots, \dot{P}_r(t)]$$

that is:

$$\mathbf{P}(t) \cdot \mathbf{A} = \dot{\mathbf{P}}(t) \tag{D.5}$$

D.4 Time dependent solution for the Markov process

To solve Equation (D.5) as a function of time we may use an analogy to ordinary differential equations in one dimension and we get:

$$\mathbf{P}(t) = \mathbf{P}(0) e^{\mathbf{A}t} \tag{D.6}$$

Although this is a very elegant solution, it is not very attractive since taking the exponential of a matrix is not that easy. Computer codes such as Matlab is required. We may, however, rewrite Equation (D.5) as:

$$\dot{\mathbf{P}}(t) = \lim_{\Delta t \to 0} \frac{\mathbf{P}(t + \Delta t) - \mathbf{P}(t)}{\Delta t} = \mathbf{P}(t) \cdot \mathbf{A}$$

yielding

$$\mathbf{P}(t + \Delta t) \approx \mathbf{P}(t)[\mathbf{A}\Delta t + \mathbf{I}] \tag{D.7}$$

where $\mathbf{I}$ is the identity matrix. This equation may now be used iteratively with a sufficient small time interval Δt and starting point $\mathbf{P}(0)$ to find the time dependent solution. Only simple matrix multiplication is required. Implementing a solution in for example VBA some considerations are required regarding the step length Δt. Choosing a too low value gives numerical problems and will also require longer computational time. Choosing a too high step length will cause the approximation in Equation (D.7) to be inaccurate. A rule of thumb will be to use a value of one tenth of the inverse value of the highest transition rate.

Note that in Markov analysis we usually only require the time-dependent solution for a limiting time period, and typically we would like to calculate $\mathbf{P}(t)$ at values $t = 0, \Delta t, 2\Delta t, \ldots$. Using Equation (D.7) is therefore attractive. To improve the approximation in Equation (D.7) we could

use one "intermediate" point, i.e., we could use:

$$\mathbf{P}(t+\Delta t) \approx \mathbf{P}(t)[\mathbf{A}\Delta t/2+\mathbf{I}][\mathbf{A}\Delta t/2+\mathbf{I}] \tag{D.8}$$

and even improve by splitting into 2^n sub-intervals, yielding:

$$\mathbf{P}(t+\Delta t) \approx \mathbf{P}(t)\left[\mathbf{A}\Delta t/2^n+\mathbf{I}\right]^{2^n} \tag{D.9}$$

Note the similarity between Equation (D.9) and Equation (11.106) in the textbook. The advantage of Equation (D.9) is the calculation efficiency, i.e., we only need n matrix multiplications to reduce the step-length by a factor 2^n. Note that we only calculate $[\mathbf{A}\Delta t/2^n+\mathbf{I}]^{2^n}$ once in Equation (D.9), so we could afford double precision in that part of the calculations to increase the precision. It should be noted that there is still a trade-off between round-off errors and accuracy in the approximation in Equation (D.9), and a good choice of n would be in the range 4-6.

Steady state solution for the Markov process

In the long run we will have that $\dot{\mathbf{P}}(t) \rightarrow \mathbf{0}$ when $t \rightarrow \infty$, hence $\mathbf{P}(t)\cdot\mathbf{A} = \mathbf{0}$. We define the steady state probabilities by the vector $\mathbf{P} = [P_1, P_2, \ldots, P_r]$, where we have omitted the time dependency (t) to reflect that in the long run the state probabilities are not changing any more.

To solve the steady state equations we realize that the matrix $\mathbf{A}$ has not full rank due to the way have have established the diagonal elements. To overcome this problem we remove one (arbitrary) of the equations in the following set of equations:

$$[P_1, P_2, \ldots, P_r]\cdot\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rr} \end{bmatrix} = [0, 0, \ldots, 0]$$

and replace it by the following equation:

$$\sum_{j=1}^{r} P_j = 1$$

For example replacing the first equation gives:

$$[P_1, P_2, \ldots, P_r] \cdot \begin{bmatrix} 1 & a_{12} & \cdots & a_{1r} \\ 1 & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & a_{r2} & \cdots & a_{rr} \end{bmatrix} = [1, 0, \ldots, 0]$$

In matrix form we write:

$$\mathbf{P} \cdot \mathbf{A}_1 = \mathbf{b} \tag{D.10}$$

where $\mathbf{b}$ is a row vector of zeros except for the first element which equals one. Note that Equation (D.10) is not on standard form $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$. Transposing each side on the equal symbol in Equation (D.10) gives $\mathbf{A}_1^{\mathrm{T}} \cdot \mathbf{P}^{\mathrm{T}} = \mathbf{b}^{\mathrm{T}}$ which could be solved by standard Gauss-Jordan elimination.

Ideally we could obtain an analytical solution for the steady state equations, but for $r > 3$ we usually stick to numerical solutions.

D.4.1 Visit frequency

The visiting frequency, ν, is one of several system performance that we define for the steady-state situation. The visiting frequency for state j, ν_j, is the unconditional transition rate into state j. We could make different arguments for the arrival rate, say ${\nu_j}^{\mathrm{arr}}$, and the departure rate, say ${\nu_j}^{\mathrm{dep}}$. Considering departures we may argue directly that:

$${\nu_j}^{\mathrm{dep}} = \alpha_j P_j \tag{D.11}$$

Similarly for arrival we have from the law of total probability:

$${\nu_j}^{\mathrm{arr}} = \sum_{k \neq j} P_k a_{kj} \tag{D.12}$$

Since in the long run we should fulfil the *balance equations* stating that the total rate into a state equals the total rate out of that state we get:

$$\nu_j = \alpha_j P_j = \sum_{k \neq j} P_k a_{kj} \tag{D.13}$$

D.5 Mean time to first passage to a given state

The visiting frequency ν_j is the unconditional transition rate into state j, whereas $1/\nu_j$ is the unconditional mean time between state j is visited. In some situation we would rather find the

mean time to the first time the system enters state j. To solve this problem we can make state j an *absorbing state*. An absorbing state means that we can not leave that state. To make a state absorbing we just remove all arcs out of that state.

Since we are considering state j as an absorbing state, we obtain the transition rate matrix identical with the original transition state matrix, except that the j'th row (corresponding to a departure) comprises only zeros. From before we know that the transition matrix has not full rank, and we may therefore remove any of the equations. It is convenient to remove the j'th column of the matrix. Further, since row j only contains zeros, $P_j(t)$ will disappear from all equations. We may therefore also remove the j'th row in the transition rate matrix. The result is a set of $r-1$ differential equations with $r-1$ unknowns, $P_1(t),\ldots,P_{j-1}(t),P_{j+1}(t),\ldots,P_r(t)$.

Note that when establishing the reduced system by removing the j'th row and the j'th column, the underlying argument is that we treat the modified system with j as an *absorbing state*, we can not do this in general. The reduced matrix is denoted $\mathbf{A}_{\mathrm{R}}$.

To solve the set of differential equations we introduce the Laplace transform. The Laplace transform of a function, say $f(t)$ is given by $f^*(s) = \mathscr{L}f(t) = \int_0^\infty e^{-st}f(t)dt$. The following rule applies for the Laplace transform:

$$\mathscr{L}[f'(t)] = s\mathscr{L}[f(t)] - f(0) = sf^*(s) - f(0) \tag{D.14}$$

In addition we have that the Laplace transform of a sum of functions equals the sum of the Laplace transforms of those functions. Now taking the Laplace transform on both sides of the set of differential equations, we observe that the right hand side is the derivative of the state probabilities, hence the Laplace of the right hand side will be $sP_i^*(s) - P_i(0)$, where $P_i(0) = 1$ only for the initial state, and 0 else.

The result is a set of $r-1$ linear equations with $r-1$ unknowns, $P_1^*(s),\ldots,P_{j-1}^*(s)$, $P_{j+1}^*(s),\ldots,$ $P_r^*(s)$. In principle we may solve these equations by elimination, or we just use the solver in our linear algebra library.

The Laplace transform of the survivor function is

$$R^*(s) = \sum_{i=1,i\neq j}^{r} P_i^*(s) \tag{D.15}$$

If we are able to take the inverse Laplace transform, we may also find the survivor function $R(t)$ of the system. A trick to do this would be to arrange the denominator on the form $(s-k_1)(s-k_2)$ and then factorize, and hope that we get something we recognize from the table of Laplace transforms of known functions.

Our objective, is however, to find the mean time to the first time the system enters state j.

We have that

$$\mathrm{E}[T] = \int_0^\infty R(t)\,dt$$

Further we also have

$$R^*(s) = \mathscr{L}R(t) = \int_0^\infty e^{-st}R(t)\,dt \tag{D.16}$$

Thus, by inserting $s = 0$ we have $\mathrm{E}(T) = \int_0^\infty R(t)e^0\,dt = R^*(0)$.

Since $R^*(0) = \sum_{i=1,i\neq j}^{r} P_i^*(0)$ we therefore obtain the mean time to first system failure by:

$$\mathrm{MTTF} = \sum_{i=1,i\neq j}^{r} P_i^*(0) \tag{D.17}$$

Note that we by this procedure may establish the mean time to the first visit to sate j without actually calculating the Laplace transforms. What we actually do is to solve a set of linear equations, where the unknown variables are the $P_i^*(0)$'s from the reduced systems by removing the row column corresponding to the absorbing state. Further note that the right hand side equals 0 for all equations except the equation representing the initial state, where the right hand side equals -1, since $sP_i^*(s) = 0$ for $s = 0$.

Note that we here have assumed that state 0 represent the system failure. In a more general setting we apply the same approach but rather than deleting the first row and column to obtain the reduced matrix, we delete the rows and columns corresponding to one or more system failure states.

D.6 Birth-death processes

A birth-death process is a special type of Markov process where the transitions are to the next state immediately above or immediately below the current state. The states has some natural ordering, for example the number of customers being served by one or more servers. For that reason we also usually start the numbering from zero rather than one. The transition matrix is then tridiagonal as shown in Equation

$$\mathbf{A} = \begin{bmatrix} a_{00} & a_{01} & 0 & \dots & \dots & \dots \\ a_{10} & a_{11} & a_{12} & 0 & \dots & \dots \\ 0 & a_{21} & a_{22} & a_{23} & 0 & \dots \\ \vdots & 0 & a_{32} & a_{33} & a_{34} & 0 \end{bmatrix} \tag{D.18}$$

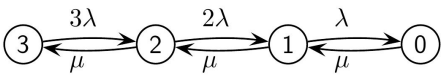


Figure D.1: Markov transition diagram

The above-diagonal elements, $a_{ij}, j-i=1$ are denoted births and causes the system state to increase by one, whereas the below-diagonal elements, $a_{ij}, i-j=1$ are denoted deaths, and causes the system state to decrease by one. In birth-death processes it is common to use λ as a transition symbol for births and μ as transition symbol for deaths. A birth-death process may have a finite or an infinite number of states.

Example

Consider a workshop with three critical machines. Each machine has a constant failure rate equal to λ and there is one repair man that can repair failed machines. The rate of repair is μ meaning that the mean repair time is $1/\mu$. The state variable represent the number of failed machines. The transition matrix is given by:

$$\mathbf{A} = \begin{bmatrix} ? & \mu & 0 & 0 \\ \lambda & ? & \mu & 0 \\ 0 & 2\lambda & ? & \mu \\ 0 & 0 & 3\lambda & ? \end{bmatrix}$$

Figure D.1 shows the Markov transition diagram corresponding to the transition matrix.

Note when the system is in state 3 and all machines are functioning, there are 3 machines that potentially may fail, hence the transition rate from state 3 to state 2 equals 3λ. In state 2 there is only two machines that may fail, hence the transition rate from state 2 to state 1 is 2λ. Since there is only one repair man, all the above-diagonal elements equal the repair rate μ.

The question marks in the transition matrix represent the diagonal elements. They are completed at the end when all the "real" transitions are specified by applying the rule that all rows should sum to one, i.e., we get:

$$\mathbf{A} = \begin{bmatrix} -\mu & \mu & 0 & 0 \\ \lambda & -\lambda-\mu & \mu & 0 \\ 0 & 2\lambda & -2\lambda-\mu & \mu \\ 0 & 0 & 3\lambda & -\lambda \end{bmatrix}$$

Figure D.2 shows the specification of this model in MS-Excel. It is convenient to give names to the cell containing λ and μ. The numerical values used are: $\lambda = 0.001$ and $\mu = 0.1$.

		To -->			
		0	1	2	3
From ↓	0		=mu	0	0
	1	=lambda		=mu	0
	2	0	=2*lambda		=mu
	3	0	0	=3*lambda	

Figure D.2: MS Excel specification of the transition matrix

Table D.1 shows the calculated steady state probabilities. Full production is achieved in 97% of the operating hours. For some 3% one machine is down for corrective maintenance, whereas the probability of two or more failed machines is very low.

Table D.1: Steady state probabilities

State	P_i
3	0.9703
2	2.91E-02
1	5.82E-04
0	5.82E-06

Problem

In a workshop there are two production lines in parallel. Each production line has a critical machine with constant failure rate $\lambda = 0.01$ failures per hour. There is one (common) spare machine that can replace a failed machine. We assume that switching time can be ignored. The repair rate of the machines is assumed constant and equal to $\mu = 0.2$ per hour. If a production line is down the loss is assumed to be $c_U = 10\ 000$ NOKs per hour. Only one repair man is available.

- Construct the Markov diagram and find the steady state solution.
- Calculate the expected loss due to downtime.
- If production is not 24/7 but runs from 07:00 to 15:00 it is reasonable to assume that each morning we start with 3 functioning machines. Find the time dependent solution and find the expected loss due to downtime.
- Repeat the analysis, but assume that two repair men are available.
- How much should one be willing to pay per hour for having this extra backup on repair resources?

D.7 Procedure

The Markov Analysis is usually carried out in six steps:

1. Make a sketch of the system
2. Define the system states
3. Group similar sates to one state (reduce dimension)
4. Draw the Markov diagram with the transition rates
5. Quantitative assessment
6. Compilation and presentation of the result from the analysis

Below we describe the state together with an example

Step 1 - Make a sketch of the system

The sketch is mainly used to visualise parallel and serial structures, stand-by systems, switching systems etc. In Figure D.3 we have drawn a sketch of a simple cold standby system. We consider a system comprising an active pump and a spare pump in cold stand-by. If the active pump fails, the stand-by pump is started and continue to do the duty. The failed active pump is then repaired. If the stand-by pump, which now is working, fails during the repair of the failed pump there will be a system failure.

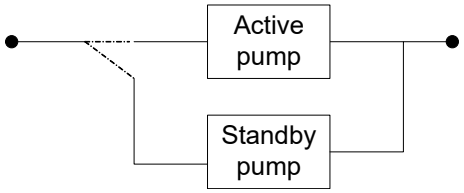


Figure D.3: Example of cold standby system

Step 2 - Define the system states

Based on the sketch of the systems the various components are identified. For each component one or more states are defined. Often a number is given to each state, where the highest number represents perfect performance, whereas zero represent a complete fault state. Next the various states of all components are combined. This may lead to very many states due to the combinatorial effect. Table D.2 shows the states for the example system.

Table D.2: States for the example system

State	Explanation
2	Active pump is running
1	Active pump failed, stand-by pump running
0	Both pumps failed

Step 3 - Group similar sates to one state (reduce dimension)

This step is only introduced in order to reduce the dimension of the problem. In many situations several components may be identical and it will usually be possible to group similar system states into one system state, and hence reduce the dimension of the problem. For example if we have $n = 3$ pumps we can group into state 3={All 3 pumps are OK}, 2={2 pumps are OK}, 1={One pump is OK} and 0={All pumps are failed}. For state 1 and 2 we do not distinguish between which pumps are functioning. In the example there is no need to group states.

Step 4 - Draw the Markov diagram with the transition rates

The various system states are now drawn in a Markov diagram. Each state is drawn as a circle labelled with the state number. Transitions between the states are visualised by drawing arrows between the corresponding circles. On each arrow the transition rate is labelled. Very often the Greek letter λ represents component failure rates, whereas the Greek letter μ represents repair rates.

Table D.3 shows the transition rates for the example system. Here we assume that both

Table D.3: Transition rates

Rate	Explanation
λ_1	failure rate of the active pump
λ_2	failure rate of the standby pump (while running, $\lambda_2 = 0$ in standby position)
μ_1	repair rate of the active pump
μ_B	repair rate when both pumps are in a fault state

pumps are repaired as part of one common repair activity if we enter state 0. Figure D.4 shows the transition diagram for the example system.

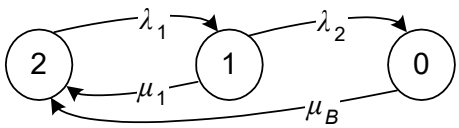


Figure D.4: Transition diagram for the example system

Step 5 - Quantitative assessment

Steady state solution:
The transition matrix **A** is given by:

$$\mathbf{A} = \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} -\mu_B & 0 & \mu_B \\ \lambda_2 & -\lambda_2 - \mu_1 & \mu_1 \\ 0 & \lambda_1 & -\lambda_1 \end{bmatrix}$$

To find the steady state solution we solve the system $\mathbf{P} \cdot \mathbf{A}_1 = \mathbf{b}$ where we may replace any column (equation) in **A** with ones. To simplify the set of equations as much as possible we now choose to replace the last column:

$$[P_0, P_1, P_2] \cdot \begin{bmatrix} -\mu_B & 0 & 1 \\ \lambda_2 & -\lambda_2 - \mu_1 & 1 \\ 0 & \lambda_1 & 1 \end{bmatrix} = [0, 0, 1]$$

To solve the set of equation we start with the second equation:

$$P_1(-\lambda_2 - \mu_1) + P_2\lambda_1 = 0 \Rightarrow P_1 = \lambda_1/(\lambda_2 + \mu_1)P_2$$

Inserted in the first equation:

$$P_0(-\mu_B) + P_1\lambda_2 = 0 \Rightarrow P_0 = \lambda_2/\mu_B P_1 = \frac{\lambda_1\lambda_2}{\mu_B(\lambda_2 + \mu_1)}P_2$$

Now P_0 and P_1 may be inserted in the third equation:

$$P_0 + P_1 + P_2 = \left[\frac{\lambda_1\lambda_2}{\mu_B(\lambda_2 + \mu_1)} + \frac{\lambda_1}{\lambda_2 + \mu_1} + 1\right] P_2 = 1$$

Multiplying with $\mu_B(\lambda_2 + \mu_1)$ on both sides and rearranging gives:

$$P_2 = \frac{\mu_B(\lambda_2 + \mu_1)}{\lambda_1(\lambda_2 + \mu_B) + \mu_B(\lambda_2 + \mu_1)}$$
$$P_1 = \frac{\mu_B\lambda_1}{\lambda_1(\lambda_2 + \mu_B) + \mu_B(\lambda_2 + \mu_1)}$$
$$P_0 = \frac{\lambda_1\lambda_2}{\lambda_1(\lambda_2 + \mu_B) + \mu_B(\lambda_2 + \mu_1)}$$

Visiting frequencies:
From $\nu_j = -P_j a_j j$ we get for example:

$$\nu_0 = -P_0 a_{00} = -P_0(-\mu_B) = \frac{\mu_B \lambda_1 \lambda_2}{\lambda_1(\lambda_2+\mu_B)+\mu_B(\lambda_2+\mu_1)}$$

Time dependent solution:
The time dependent solution requires to solve the Laplace equations, and is rather complicated. Therefore we stick to numerical methods. At the end of this document we demonstrate the use of Laplace to find the time dependent solution for a simpler situation with only one component.

Mean time to first system failure:
We use the Laplace transform approach. That is, first we delete the row and column corresponding to the absorbing state, i.e., state 0, and replace the P_j's with P_j^*s:

$$[P_1^*, P_2^*] \cdot \begin{bmatrix} -\lambda_2-\mu_1 & \mu_1 \\ \lambda_1 & -\lambda_1 \end{bmatrix} = [0, -1]$$

with the solution $P_1^* = 1/\lambda_2$ and $P_2^* = (\lambda_2+\mu_1)/(\lambda_1\lambda_2)$, and thus:

$$\text{MTTF}_S = P_1^* + P_2^* = (\lambda_1+\lambda_2+\mu_1)/(\lambda_1\lambda_2)$$

D.8 Time dependent solution for a repairable component

Consider a component with constant failure rate λ and constant repair rate μ. Let state 1 represent the functioning state and state 0 represent the failed state. The transition matrix for this system is given by:

$$\mathbf{A} = \begin{bmatrix} -\mu & \mu \\ \lambda & -\lambda \end{bmatrix}$$

Assuming the system starts in state 1 we have $P_0(0) = 0$ and $P_1(0) = 1$, and the Laplace transform of the time dependent solution is given by:

$$[P_0^*(s), P_1^*(s)] \begin{bmatrix} -\mu & \mu \\ \lambda & -\lambda \end{bmatrix} = [sP_0^*(s), sP_1^*(s) - 1]$$

Thus

$$-\mu P_0^*(s) + \lambda P_1^*(s) = sP_0^*(s)$$
$$\mu P_0^*(s) - \lambda P_1^*(s) = sP_1^*(s) - 1$$

Adding these two equations yields:

$$sP_0^*(s) + sP_1^*(s) = 1 \Rightarrow P_0^*(s) = 1/s - P_1^*(s)$$

and inserting into the last of the above equations:

$$\mu/s - \mu P_1^*(s) - \lambda P_1^*(s) = sP_1^*(s) - 1$$

which is solved wrt $P_1^*(s)$:

$$P_1^*(s) = \frac{1}{\lambda+\mu+s} + \frac{\mu}{s}\frac{1}{\lambda+\mu+s}$$

This expression is not recognized in the list of Laplace transforms. A trick is now to multiply with $(\lambda+\mu)/(\lambda+\mu)$:

$$P_1^*(s) = \frac{\lambda+\mu}{\lambda+\mu}\left(\frac{1}{\lambda+\mu+s} + \frac{\mu}{s}\frac{1}{\lambda+\mu+s}\right) =$$
$$\frac{\lambda}{\lambda+\mu}\cdot\frac{1}{\lambda+\mu+s} + \frac{\lambda}{\lambda+\mu}\cdot\frac{\mu}{s}\cdot\frac{1}{\lambda+\mu+s} + \frac{\mu}{\lambda+\mu}\cdot\frac{1}{\lambda+\mu+s} + \frac{\mu}{\lambda+\mu}\cdot\frac{\mu}{s}\cdot\frac{1}{\lambda+\mu+s}$$
$$= \frac{\lambda}{\lambda+\mu}\cdot\frac{1}{\lambda+\mu+s} + \frac{\mu}{s}\cdot\frac{\lambda+s+\mu}{\lambda+\mu}\cdot\frac{1}{\lambda+\mu+s} = \frac{\lambda}{\lambda+\mu}\cdot\frac{1}{\lambda+\mu+s} + \frac{\mu}{\lambda+\mu}\cdot\frac{1}{s}$$

Now using $\mathscr{L}e^{\alpha t} = 1/(s-\alpha)$ and $\mathscr{L}1 = 1/s$ we find the inverse Laplace of $P_1^*(s)$ $(-\alpha = \lambda+\mu)$:

$$d$$

and

$$P_0(t) = 1 - P_0(t) = \frac{\lambda}{\lambda+\mu}\left(1 - e^{-(\lambda+\mu)t}\right)$$

Note that when $t > 3/(\lambda+\mu)$ the time dependent solution is deviating from the steady state solution with only 5%. In practice, we therefore often say that steady state is achieved after 3 times the shortest expected transition time, here $3/\mu$. The time dependent solution is often needed in FTA (fault tree analysis) and RBD (reliability block diagram) analyses.

Appendix E

Calculating Q_0 in the PF-model

In this Appendix we will describe the method used for calculating the probability that a potential failure is not detected by the inspection regime. There are two main sources for not detecting a potential failure in due time; *i*) the inspection interval is to long compared to the PF- interval, and *ii*) the quality of the inspection is too low to detect a potential failure. The following quantities are defined:

- T_{PF} = PF interval (stochastic variable).
- $\xi(t)$ = Probability density function of T_{PF}
- q = Failure probability of one inspection
- q_{C} = Common cause part of q
- q_{I} = Independent part of q
- τ = Inspection interval

The probability that the inspection strategy fails to reveal a potential failure before a critical failure occurs could be found by the low of total probability:

$$Q_0(\tau,\xi,q) = \int_0^\infty Q_t(\tau,q,t)\xi(t)\,dt \tag{E.1}$$

where $Q_t(\tau,q,t)$ is the probability of not detecting a potential failure given that the PF interval, T_{PF}, equals t. In order to calculate $Q_t(\tau,q,t)$ we observe that when $T_{\mathrm{PF}} = t$, then number of possibilities to detect a failure equals n or $n+1$ where $n = \mathrm{int}(t/\tau)$ and $\mathrm{int}(\cdot)$ is the integer function.

The probability that we will have $n+1$ opportunities equals $t/\tau - n$ and thus the probability that we will have n opportunities to detect a potential failure equals $n+1-t/\tau$. Since the probability that a given inspection fails to detect a potential failure equals q, $Q_t(\tau,q,t)$ could easily

be obtained by:

$$Q_t(\tau, q, t) = (n + 1 - t/\tau)q^n + (t/\tau - n)q^{(n+1)} \tag{E.2}$$

if the inspections could be considered statistically independent. However, the assumption that inspections are independent does not seem realistic. A more realistic assumption would be to assume that the failure probability of one inspection is given by:

$$q = q_C + q_I \tag{E.3}$$

where q_C represents common cause failures due to systematic failures such as low coverage, and q_I represents the failure probability due to specific conditions for one run, e.g., inadequate velocity of the measuring wagon, human errors etc.

Assuming that the failure probability of one inspection could be split into a common and an independent part as shown in Equation (E.3) we calculate the total failure probability of the inspection strategy as:

$$Q_0^*(\tau, \xi, q_C, q_I) = 1 - (1 - q_C)\left[1 - Q_0(\tau, \xi, q_I)\right] \tag{E.4}$$

Codes for implementing $Q_0(\tau, \xi, q)$ in this course assume that $\xi(t)$ is the gamma probability density function with some expected value and standard deviation. If we use the gamma distribution the format of the Q_0-function will be $Q_0(\tau, E_{PF}, SD_{PF}, q_I)$.

Bibliography

Backer, R. and Christer, A. (1994). Review of delay-time or modelling of engineering aspects of maintanannce. *European Journal of Operation Research*, 73:407–442.

Badihi, H., Zhang, Y., Jiang, B., Pillay, P., and Rakheja, S. (2022). A Comprehensive Review on Signal-Based and Model-Based Condition Monitoring of Wind Turbines: Fault Diagnosis and Lifetime Prognosis. *PROCEEDINGS OF THE IEEE*, 110(6):754–806.

Bladt, M. and Sørensen, M. (2009). Efficient estimation of transition rates between credit ratings from observations at discrete time points. *Quantitative Finance*, 9(2):147–160.

Boersma, S., Doekemeijer, B., Gebraad, P., Fleming, P., Annoni, J., Scholbrock, A., Frederik, J., and van Wingerden, J.-W. (2017). A tutorial on control-oriented modeling and control of wind farms. In *2017 American Control Conference (ACC)*, pages 1–18.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B.*, 34 (2):187–220.

DNV-RP-A204 (2021). Qualification and assurance of digital twins - recommended practice. Standard, DNV, Høvik, NO.

DNV-RP-G101 (2021). Risk based inspection of offshore topsides static mechanical equipment - recommended practice. Standard, DNV, Høvik, NO.

Fox, H., Pillai, A. C., Friedrich, D., Collu, M., Dawood, T., and Johanning, L. (2022). A Review of Predictive and Prescriptive Offshore Wind Farm Operation and Maintenance. *Energies*, 15(2):504.

Gartner (2019). Top 10 strategic technology trends for 2019. https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2019.

Göçmen, T., van der Laan, P., Réthoré, P.-E., Diaz, A. P., Larsen, G. C., and Ott, S. (2016). Wind turbine wake models developed at the technical university of denmark: A review. *Renewable and Sustainable Energy Reviews*, 60:752–769.

Howland, M. F., Bossuyt, J., Martínez-Tossas, L. A., Meyers, J., and Meneveau, C. (2016). Wake structure in actuator disk models of wind turbines in yaw under uniform inflow conditions. *Journal of Renewable and Sustainable Energy*, 8(4):043301.

ISO14224 (2016). Petroleum, petrochemical and natural gas industries - collection and exchange of reliability and maintenance data for equipment. Standard, International Organization for Standardization, Geneva, CH.

Jensen, N. O. (1983). Note on wind generator interaction. (wakes). Technical Report RISO-M-2411.

Jimenez, A., Crespo, A., and Migoya, E. (2009). Application of a les technique to characterize the wake deflection of a wind turbine in yaw. *Wind Energy*, 13(6):559–572.

Katic, I., Højstrup, J., N.O., and Jensen (1986). A simple model for cluster efficiency. In Palz, W. and Sesto, E., editors, *EWEC'86. Proceedings. Vol. 1*, pages 407–410. A. Raguzzi.

Kawauchi, Y. and Rausand, M. (1999). Life cycle cost (lcc) analysis in oil and chemical process industries. Technical Report NTNU.

Kirwan, B. and Ainsworth, L. K. (1997). *A Guide to Task Analysis*. Taylor & Francis, London.

Kwang Pil, C., Rausand, M., and Vatn, J. (2008). Reliability assessment of reliquefaction systems on lng carriers. *Reliability Engineering & System Safety*, 93(9):1345–1353. Safety in Maritime Transportation.

Laskowska, E. and Vatn, J. (2020). Degradation modelling using a phase type distribution (phd). In *Proceedings of the 30th European Safety and Reliability Conference(ESREL)*, pages 3569–3576.

Laskowska, E. M., Vatn, J., and Yin, S. (2023). Prognostic model for reliability assessment of emergency shutdown valves used in oil & gas facility. *International Journal of Reliability and Safety*. In press.

Lee, J. and Zhao, F. (2022). GWEC | Global Wind Report 2022. Technical report.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Moubray, J. (1991). *Reliability-centred Maintenance*. Butterworth-Heinemann, Oxford.

Nowlan, F. S. and Heap, H. F. (1978). Reliability-centered maintenance. Technical report, National Technical Information Service, US Department of Commerce, Springfield, Virginia, Technical Report AD/A066-579.

Øien, K., Hokstad, P., and Rosnes, R. (1998). Life cycle cost (lcc) analysis in oil and chemical process industries. Technical Report - STF38 A98419.

Pedersen, V. (2020). Cmms - computerized maintenance management systems, lecture notes. Technical report.

Rausand, M., Barros, A., and Høyland, A. (2021). *System Reliability Theory: Models, Statistical Methods, and Applications*. John Wiley & Sons, Inc.

Shakoor, R., Hassan, M. Y., Raheem, A., and Wu, Y.-K. (2016). Wake effect modeling: A review of wind farm layout optimization using jensen's model. *Renewable and Sustainable Energy Reviews*, 58:1048–1059.

Technologies, W. F. . (2022). Wind - Fuels & Technologies.

Thiruthiyappan, T. (2022). Predictive maintenance for offshore wind farms - specialization project, ntnu. Technical report.

Vatn, J. (2023). Optimizing ultrasonic inspection regimes of railway rails. In *Proceedings of the 33rd European Safety and Reliability Conference(ESREL)*.

Wintle, J. B., Kenzie, B., Amplett, G., and Smalley, S. (2001). Best practice for risk based inspection as a part of plant integrity management. Standard, HSE, ISBN 0 7176 2090 5.

Zhang, W., Vatn, J., and Rasheed, A. (2022). A review of failure prognostics for predictive maintenance of offshore wind turbines. *Journal of Physics: Conference Series*, 2362(1):012043. Publisher: IOP Publishing.