Evaluation of HYDRA

A risk model for hydropower plants

LARS ALMGREN
Evaluation of HYDRA -
A risk model for hydropower plants

LARS ALMGREN

Master’s Thesis in Financial Mathematics (30 ECTS credits)
Master Programme in Industrial Engineering and Management (120 credits)
Royal Institute of Technology year 2016
Supervisor at Vattenfall Hydro AB:
Hanna Alsén, Fredrik Engström, Mattias Androils
Supervisor at KTH: Boualem Djehiche
Examiner: Boualem Djehiche

TRITA-MAT-E 2016:71
ISRN-KTH/MAT/E--16/71--SE

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci
Abstract

Vattenfall Hydro AB has more than 50 large scale power plants. In these power plants there are over 130 power generating units. The planning of renewals of these units is important to minimize the risk of having big breakdowns which inflict long downtime. Because all power plants are different Vattenfall Hydro AB started using a self developed risk model in 2003 to improve the comparisons between power plants. Since then the model has been used without larger improvements or validation.

The purpose of this study is to evaluate and analyse how well the risk model has performed and is performing. This thesis is divided into five subsections where analyses are made on the input to the model, adverse events used in the model, the probabilities used in the model, risk forecasts from the model and finally trends for the periods the model has been used. In each subsection different statistical methods are used for the analyses.

From the analyses it is clear that the low number of adverse events in power plants makes the usage of statistical methods for evaluating performance of Vattenfall Hydro AB’s risk model imprecise. Based on the results of this thesis the conclusion is made that if the risk model is to be used in the future it needs further improvements to generate more accurate results.
Sammanfattning

Vattenfall Vattenkraft AB har fler än 50 storskaliga vattenkraftverk. I dessa finns det totalt över 130 stycken aggregat som omvandlar energi. För att minimera risken för stora avbrott är planeringen av förnyelser av dessa aggregat viktigt. 2003 började Vattenfall Vattenkraft AB använda en egenutvecklad riskmodell för att lättare kunna jämföra riskerna mellan kraftverken. Sedan dess har man använt modellen utan större förbättringar och validering.

Syftet med detta examensarbete är att utvärdera och analysera hur väl riskmodellen fungerar och har fungerat. Studien är uppdelad i fem sektioner där analyser görs på indata till modellen, oönskade händelser som används i modellen, sannolikheter som används i modellen, riskprognoser från modellen och slutligen trender för perioden då modellen använts.

Baserat på resultaten från denna studie är slutsatsen att om riskmodellen ska fortsätta användas i framtiden behöver man göra förbättringar för att få fram mer precisa resultat. Från resultaten står det även klart att det låga antalet oönskade händelser i kraftverken påverkar de statistiska modeller som används för analysera riskmodellen.
Acknowledgements

I would like to thank Vattenfall Hydro AB for the opportunity to write my master thesis with them. Especially I would like to thank my supervisors Hanna Alsén, Fredrik Engström and Mattias Androls. I also want to thank Boualem Djehiche, my supervisor at KTH, for all the help he has given me.

Stockholm, October 2016
Lars Almgren
Contents

1 Introduction .......................... 1
  1.1 Background .......................... 1
    1.1.1 Hydropower ....................... 1
    1.1.2 Asset management .................. 2
  1.2 Purpose ........................... 3
  1.3 Problem statement .................... 3
  1.4 Limitation .......................... 4
    1.4.1 Imperfect data ...................... 4
  1.5 Delimitations ....................... 5
    1.5.1 Generators and turbines .......... 5
    1.5.2 Dams ............................. 5
  1.6 Method .............................. 5
    1.6.1 Availability ....................... 6
    1.6.2 Expenditure ....................... 6
    1.6.3 Failures ........................... 6
    1.6.4 Insurance events ................... 6
    1.6.5 Lifetime ........................... 6
    1.6.6 Lost margin ......................... 7
    1.6.7 Outage ............................ 7
    1.6.8 Plant attributes and condition .... 7
    1.6.9 Work spill .......................... 7

2 Theoretical framework .......... 9
  2.1 Definition of risk .................... 9
  2.2 The risk model ....................... 9
    2.2.1 Hydropower plant risks .......... 10
    2.2.2 Condition of plant parts .......... 10
    2.2.3 Plant attributes ................... 10
    2.2.4 The probabilities of adverse events .... 11
  2.3 Kruskal-Wallis test ................... 11
  2.4 Spearman test ........................ 12
  2.5 Reliability analysis .................. 13
    2.5.1 The survival function ............ 13
2.5.2 Right-censoring ........................................ 14
2.5.3 Non-parametric estimators ............................... 16
  2.5.3.1 Kaplan-Meier estimator .......................... 17
2.5.4 Log-log confidence intervals .......................... 18
2.5.5 Log-rank test ........................................... 19
2.6 Cox Proportional Hazards model .......................... 20
  2.6.1 Regression analysis .................................... 20
  2.6.2 The model .............................................. 21
  2.6.3 Fitting the proportional hazards regression model ..... 22
2.6.4 Verification of the proportional hazards assumption ...... 23
  2.6.5 Model validation ....................................... 24
2.7 Aalen’s additive model .................................... 25

3 Analysis ................................................................ 29
  3.1 Condition of the plants ...................................... 29
  3.2 Adverse events ............................................... 31
  3.3 Probabilities .................................................. 32
  3.4 Risk forecasts ............................................... 34
  3.5 Consequences and trends .................................... 34

4 Results .................................................................. 37
  4.1 Condition of the plants ...................................... 37
    4.1.1 Condition against the availability .................... 37
    4.1.2 Condition against hours of downtime due to production-related interruptions ......................... 38
  4.2 Adverse events ............................................... 40
    4.2.1 Dataset 1 ................................................ 40
      4.2.1.1 Case 1 ............................................. 40
      4.2.1.2 Case 2 ............................................. 40
    4.2.2 Dataset 2 ................................................ 41
      4.2.2.1 Case 1 ............................................. 42
      4.2.2.2 Case 2 ............................................. 43
  4.3 Probabilities .................................................. 44
    4.3.1 Kaplan-Meier estimator ............................... 44
    4.3.2 Two regression models .................................. 48
      4.3.2.1 Aalen’s additive model ......................... 49
      4.3.2.2 The Cox model .................................. 51
    4.3.3 Comparisons of the probabilities ..................... 53
  4.4 Risk forecasts ............................................... 54
    4.4.1 Boden 2013 ............................................. 54
    4.4.2 Juktan 2011 ............................................. 55
    4.4.3 Olidan 2010 ............................................. 55
    4.4.4 Torpshammar 2009 .................................... 56
  4.5 Consequences and trends .................................... 57
Chapter 1

Introduction

1.1 Background

In this section we give a brief introduction to the management of hydropower. Having some knowledge of these fields will make it easier to understand certain technical terms and concepts.

1.1.1 Hydropower

To ensure a continuous and reliable delivery of energy to the society Vattenfall AB, further on denoted as Vattenfall, produces heat and electricity from several different sources. Heat and electricity are produced from six different energy sources: wind power, hydropower, biomass, nuclear power, coal and gas. Vattenfall focuses on improvement of sustainable solutions and production with low emissions.

As a renewable energy source with low risk of environmental hazards, hydropower plays an important role in Vattenfall’s energy production. In fact, hydropower is also of great importance for the Swedish energy market where every year it produces around 45 percent of all electricity. Since the frequency in the power grid constantly has to be 50 hz, another positive feature of hydropower is its controllability, where it is easy to increase or decrease the production. [1]

Within hydropower energy is extracted from flowing water. Figure 1.1 shows a sketch of a typical hydropower plant. Water is led from a reservoir to a turbine. The water then flows through the turbine, where the turbine uses the flow of the water to create mechanical energy which makes the turbine spin. The mechanical energy is transferred to a generator that generates electric energy. Lastly the electricity goes through a transformer which transforms
it to a certain voltage and the electricity is then distributed to households and industries. [2]

Within hydropower it is common practise to refer to the generator and the turbine together as a unit and hence this was done in this thesis as well.

1.1.2 Asset management

To understand what asset management is in the context of energy it is first important to understand what is meant by an asset. An asset could be financial, physical, human, information or intangible [4]. In this thesis physical assets will be the only assets to be considered, namely hydropower plants. One definition of assets in engineering organizations is

\[\text{Physical components of manufacturing, production or service facility, which has value, enables services to be provided, and has an economic life greater than twelve months} \] [5].

Asset management within this thesis is thus the management of these components. Haider [5] gives the definition of asset management as a set of activities associated with:

- Identifying what assets are needed
- Identifying funding requirements
- Acquiring assets
- Providing logistic and maintenance supports systems for assets
• Disposing or renewing assets

This is done to effectively and efficiently meet the desired objective.

If instead the work of an asset manager is considered Balzer et al. summarize it well with a number of steps to be carried out [6].

1. Determining the overall strategy for all pieces of equipment.
2. Implementation of the overall strategy and derivation of particular asset decisions on equipment level.
3. Selection of the appropriate maintenance activity.
4. Optimal maintenance activity of the asset.

Hence, asset management and the work of an asset manager can be summarized as making sure that production is fluent and that the risk of failures are balanced. [7]

1.2 Purpose

The purpose of this study is to investigate if the risk model used today deliver the expected risk assessments, and to analyze data that has been collected during the last ten years to see how well the results from the model predicts the risks.

1.3 Problem statement

During the last year the spot price of electricity has been historically low, mainly because of the increasing number of wind power plants [8]. Add to this transmission fees and a property tax for hydropower plants (about 0.09 Swedish SEK / kWh [9]). It is easy to understand that the margin between the sport price and the costs are small. Due to this it is of great importance to optimize operations within Vattenfall.

The first thing that probably comes to mind is finding a method for optimizing the production. This is however no the only way to generate better results. An equally important method is the optimization of maintenance and upgrading of power plants. With good planning it is possible to keep the costs low and at the same time prevent breakdowns.

At Vattenfall Hydro the Asset Management Department is responsible for the investment planning of the power plants. To support the investment decisions the asset managers at Vattenfall Hydro uses a self-developed risk
model to forecast risks in their plants. The model they use today has been used for about 13 years. However Vattenfall Hydro does not really know how well the model performs and therefore they do not know how reliable the model’s forecasts are. Hence, there has been a need to investigate if the model predicts, and have been predicting for the period it has been used, risks in a correct manner.

This thesis intends to answer the question:

How reliant is the risk model Vattenfall Hydro is using today to forecast the risks within their hydropower plants for the coming year?

This will be done by investigating five delimited questions:

- How well can Vattenfall Hydro assess the condition of the power plants?
- Has Vattenfall Hydro chosen the right set of adverse events for the model?
- Do the probabilities in the risk model reflect reality in a correct manner?
- Does the model successfully predict the risks in the power plants?
- What have the consequences and trends been during the period?

### 1.4 Limitation

#### 1.4.1 Imperfect data

The dataset used in this thesis is not perfect. The way some of the datasets are structured made it hard to use proper statistical methods to evaluate the quality of the risk assessments of the risk model. Due to this fact, the
results from the evaluation of the assessments of the condition of the power plants should be interpreted with this in mind.

1.5 Delimitations

1.5.1 Generators and turbines

The risk model is based on probabilities and consequences of adverse events. Adverse events can be understood as events in a power plant that interrupt the production, i.e. fire in a generator. The adverse events were chosen in the process when the model was created and have been the same ever since. In the list of adverse events there are more than 40 kinds, such as turbine servo breakdown, flooding, and leakage. Because of the big amount of events this thesis focuses on adverse events for the generators and turbines. This is done because, except for the dam, generators and turbines are the two most important components of a hydropower plant.

1.5.2 Dams

This thesis does not cover any risks considering dams. The decision to leave these risks out was mainly based on the fact that this risk model does not consider risks of dams in any large scale. Another reason is that it is hard to estimate the consequences and probabilities of dam breakdowns. However, it should be clarified that dam risks are not overlooked outside the scope of this thesis. Vattenfall has a designated division for assessing only risks of dams.

1.6 Method

In this thesis a quantitative analysis is used by analyzing data that have been collected during the lifetime of the risk model. There are also analyses of the quality of the engineering assessments by comparing these with collected data during the same period.

For the analysis several datasets have been used. Below follows a description of each dataset and an estimate of its quality.
1.6.1 Availability

The dataset of availability is obtained as availability per unit per power plant per year. There are two types of availability, one with planned downtime and one without. This data was collected during the period from 2005 until 2015. It has been collected daily and can be considered to be of excellent quality.

1.6.2 Expenditure

The dataset of expenditure is obtained as type of yearly expenditure per power plant. The types of expenditure are operational expenses (OPEX) and capital expenditures (CAPEX). The data was assessed during the period from 2006 until 2015. The data can be considered to be of good quality.

1.6.3 Failures

There are two datasets of failures, one with all types of failures in the power plants and one to specified production-related failures. The dataset with specified production-related failures has records for the period from 2012 until 2015. The dataset with all types of failures has records from the time period 2004 until 2015. In both datasets the data is obtained as type of failures, power plant, unit, the part that is related to the failure, and date of failure. These datasets can be considered to be of fair quality for the type of analyses they will be used in.

1.6.4 Insurance events

The dataset of insurance events is obtained as type of event, damaged part, date of event and estimated cost. This data can be considered to be of good quality.

1.6.5 Lifetime

There are two datasets of lifetime, one for generators and one for turbines. The dataset of generators is divided into plant, effect, year of manufacturing and renewal/improvement of quality. The dataset of turbines is divided into plant, type of turbine, year of manufacturing and renewal/improvement of quality. These datasets can be considered to be of fair quality.
1.6.6 Lost margin

When a hydropower plant cannot use the water in the dam for some reason, it sometimes has to spill water to not exceed limits of the dam. This spill can be seen as a loss of potential energy conversion and in extension a loss of income. This loss of income is the lost margin. It is defined as the amount of water that is spilled times the energy conversion quotient of the hydropower plant times the spot price. The lost margin has thus a direct relation to the work spill. This dataset is obtained from calculations with data from the datasets work spill (see further down) and outage (bellow).

1.6.7 Outage

The dataset of outage is obtained as type of outage per generator per hour for a total of 129 units. The data was collected during the period from 2005 until 2015. This data is collected daily and is considered to be of excellent quality.

1.6.8 Plant attributes and condition

The dataset of plant attributes and conditions is obtained as engineering assessments once every year. This data was collected during the period from 2004 until 2015. Part of this thesis is to assess the quality of this data.

1.6.9 Work spill

Work spill is the amount of water that is spilled from a hydropower plant due to some interrupt of one or more units in the plant. The dataset of work spill is obtained as work spill per power plant per hour for a total of 55 hydropower plants. The data was collected during the period from 2005 until 2015. The data is collected daily and is considered to be of excellent quality.
Chapter 2

Theoretical framework

2.1 Definition of risk

The definition of risk $R$ used in this thesis is the product of the probability $P$ of an event and a measure of the consequence $c$ from the event, equation 2.1 [10]. These are probabilities of adverse events and the costs of repair and production downtime due to adverse events.

$$ R = P \cdot c $$

2.2 The risk model

The risk model is named HYDRA. It calculates the aggregated risk from a list of predetermined adverse events. Let $n$ be the number of adverse events and denote the $j$:th adverse event as $E_j$. Examples of adverse events could be stator breakdown or fire in the power plant.

$$ E_j = \text{Stator breakdown} $$

Each event has a probability to occur depending on the condition of parts in the power plants, units and transformers (if distinction is not necessary, power plants, units and transformers will hereby be referred to as risk objects). The probability of an event is also dependent of attributes of the risk objects. An attribute for a turbine is for example what type of turbine it is, and for a generator an attribute is the effect of the generator. To each event there is also a deterministic cost of repair that is dependent of the attributes of the risk objects. Due to the adverse events there is also a cost of the lost production. For each event some combinations of attributes and conditions have higher risks than other.
2.2.1  Hydropower plant risks

HYDRA calculates four different aggregated risks for each hydropower plant, the risk of the units, the transformers, the switchgear, and an overall risk of each power plant. An aggregated risk is calculated as the sum of the risks of a set of predetermined adverse events.

\[ R_i = \sum_{j=1}^{n} p_{ij} c_j \]  

\( R_i \) - Aggregated risk of plant \( i \)  
\( n \) - The number of adverse events  
\( p_{ij} \) - Probability of adverse event \( j \) for plant \( i \)  
\( c_j \) - Cost of repair and production loss due to adverse event \( j \)

2.2.2  Condition of plant parts

The risk objects consist of a number of different parts that are related to the adverse events. The condition of each part is estimated by the maintenance staff, technical specialists and asset managers. Each part \( k \) can either have the condition \( S_k \) "good" or "bad".

\[ S_k = \begin{cases} 1 & \text{if condition is good} \\ 0 & \text{if condition is bad} \end{cases} \]  

2.2.3  Plant attributes

The risk objects have specific attributes that influence the risk. Examples of attributes could be if the turbine of a unit is of Kaplan or Francis type or if the winding of the stator is shellac, asphalt or epoxy. Different combinations of these attributes give different risk in HYDRA by influencing the probability and the consequences of the event.

Let \( A_l \) denote an attribute, where \( l \) denotes which attribute that is considered, i.e. \( A_{\text{Turbine type}} \). For each attribute there is at least two possible choices but there could be even more. The number of possible choice for attributes is denoted as \( m \). In the case of which turbine type the choices are Kaplan or Francis and thus the relationship is

\[ A_{\text{Turbine type}} = \begin{cases} \text{Kaplan} & \\ \text{Francis} & \end{cases} \]  

10
and more generally

\[ A_l = \begin{cases} 
\text{Choice 1} \\
\text{Choice 2} \\
\vdots \\
\text{Choice } m
\end{cases} \quad (2.6) \]

### 2.2.4 The probabilities of adverse events

The probability of an adverse event, Equation 2.7, is dependent of the condition of parts related to the event and the attributes of parts related to the event. Here parts related to the event could be for example the stator for a stator hazard or the runner for a runner hazard.

\[ p_{ij} = P(E_j \mid S_1, S_2, \ldots, S_n, A_1, A_2, \ldots, A_m) \quad (2.7) \]

- \( p_{ij} \) - Probability of adverse event \( j \) for plant \( i \)
- \( E_j \) - Adverse event \( j \)
- \( S_k \) - Condition of part \( k \), \( k = 1, \ldots, n \)
- \( m \) - The number of parts
- \( A_l \) - Attribute \( l \), \( l = 1, \ldots, m \)
- \( m \) - The number of possible choices of condition \( l \)

### 2.3 Kruskal-Wallis test

The Kruskal-Wallis test is used to test if hypothesis data are supposed to come from the same distribution against the alternative that they are from different distributions. The test uses ranks instead of the original values \( X_{ji} \) and for large samples sizes it approximately follows the \( \chi^2 \) distribution with \( C - 1 \) degrees of freedom, where \( C \) are the number of groups.

Assume that \( X_1 = (X_{11}, \ldots, X_{1n_1})^T, \ldots, (X_{C1}, \ldots, X_{Cn_C})^T \) are samples from \( C \) independent populations with absolutely continuous cumulative distribution functions (CDFs) \( F_1(x), \ldots, F_C(x) \). Denote the rank of \( X_{ji} \) in the pooled sample as

\[ R_{ji} = \sum_{r=1}^{C} \sum_{u=1}^{n_r} (X_{ru} \leq X_{ji}), \quad (2.8) \]

and

\[ R_j = \sum_{i=1}^{n_j} R_{ji} \quad (2.9) \]
where \( j = 1, \ldots, C \). The test statistic of Kruskal-Wallis, \( F_{KW} \), is defined as

\[
F_{KW} = \frac{12}{n(n+1)} \sum_{j=1}^{C} \frac{1}{n_j} \left( R_j - n_j \frac{n+1}{2} \right)^2
\]

(2.10)

where \( n = \sum_j n_j \) and \( \frac{n+1}{2} = \sum_j \frac{R_{i,j}}{n} \).

For small values of \( F_{KW} \) the hypothesis \( H_0 \) holds and for large values, \( F_{KM} > f_{KW, \alpha} \) the null hypothesis should be rejected at the level \( \alpha \) of significance. If the sample size is large the asymptotic \( \chi^2_{C-1, \alpha} \)-distribution can be used. It is also possible to use the \( p \)-value rejection rule, where the \( p \)-value represents the probability that the test statistics in \( H_0 \) assumes greater or equal value than the observed one and comparing it with \( \alpha \).

In the presence of tied observations, the average of ranks corresponding to the position value the tied observations cover should be assigned to the observations in tied-groups. This can be used in the standard procedure and to the large-sample approximation. [11]

### 2.4 Spearman test

Spearman’s \( \rho \) is a measure of correlation. Spearman’s \( \rho \) has a distribution function that is independent of the distribution function of \( (X, Y) \) if \( X \) and \( Y \) are independent and continuous. Hence this measure can be used as a test statistic to non-parametrically test the null hypothesis that \( X \) and \( Y \) are independent. This test may also be used for ordered categorical data.

Let \( \bar{R}(X_i) \) be the rank of \( X_i \) compared to the other \( X \)-values for \( i = 1, \ldots, n \) and similarly for \( Y \). If there are ties, the mid rank is assigned to each tie, which is the average of the ranks that would have been assigned if there had been no ties. The correlation coefficient is then

\[
\rho = \frac{\sum_{i=1}^{n} (\bar{R}(X_i) - \bar{R}(X)) (\bar{R}(Y_i) - \bar{R}(Y))}{\left[ \sum_{i=1}^{n} (\bar{R}(X_i) - \bar{R}(X))^2 \sum_{i=1}^{n} (\bar{R}(Y_i) - \bar{R}(Y))^2 \right]^{1/2}}
\]

\[
= \frac{\sum_{i=1}^{n} \bar{R}(X_i) \bar{R}(Y_i) - n \bar{R}(X) \bar{R}(Y)}{\left( \sum_{i=1}^{n} (\bar{R}(X_i))^2 - n (\bar{R}(X))^2 \right)^{1/2} \left( \sum_{i=1}^{n} (\bar{R}(Y_i))^2 - n (\bar{R}(Y))^2 \right)^{1/2}},
\]

where \( \bar{R}(X) = \bar{R}(Y) = \frac{n+1}{2} \) denotes the average rank of \( X \) and \( Y \).

To test independence let \( F_{X,Y} \) denote the joint distribution function of the bivariate random variable \( (X, Y) \), and let \( F_X \) and \( F_Y \) denote the marginal
distributions of $X$ and $Y$, respectively. The null hypothesis is then that $X$ and $Y$ are mutually independent,

$$H_0 : F_{X,Y}(x, y) = F_X(x)F_Y(y), \quad \forall (x, y) \in \mathbb{R}^2.$$ 

The alternative hypotheses corresponds to the concepts of positive and negative correlation between $X$ and $Y$.

(i) One-sided test for positive correlation,

$$H_1 : X \text{ and } Y \text{ are positively correlated.}$$

(ii) One-sided test for negative correlation,

$$H_1 : X \text{ and } Y \text{ are negatively correlated.}$$

For both tests the test statistics is

$$S = \sum_{i=1}^{n} [(R(X_i) - R(Y_i))^2]. \quad (2.11)$$

The $p$-value for each test will differ according to the alternative hypothesis of interest. In case (i) the $p$-value is the probability that $S$ is less than or equal to its observed value. Thus a $p$-value which is smaller than some confidence level $\alpha$ imply there is positive correlation. In case (ii) the p-value is 1 minus the p-value of case (i) since it is the opposite case. Thus a $p$-value which is smaller than some confidence level $\alpha$ imply there is negative correlation. [11]

### 2.5 Reliability analysis

Reliability analysis is a method for determining the expected time duration until an event is likely to happen [12]. It is a good analytical method to use when estimating the lifetime of technical systems, such as generators or turbines.

#### 2.5.1 The survival function

Let $T \geq 0$ be a random variable representing the time to an event of interest. Probability theory says that the probability that $T \leq t$ is equal to the Cumulative Distribution Function (CDF)

$$P(T \leq t) = F_T(t). \quad (2.12)$$
Since $T$ is the only random variable the subscript of the CDF can be dropped. From the CDF the *Probability Density Function (PDF)* can be obtained by

$$f(t) = \frac{d}{dt} F(t). \quad (2.13)$$

Define a function $S(t)$ that is the complement to the CDF,

$$S(t) = 1 - F(t) = P(T > t), \quad 0 < t < \infty \quad (2.14)$$

Function 2.14 is known as the *Survival Function*. This function defines the probability of surviving up to a time $t$. The Survival function takes values between 0 for $t \to \infty$ and 1 for $t = 0$. [13]

The Survival function can be defined by another function called the *Hazard Function*. The Hazard Function is the probability, given that the event has not occurred up to time $t$, that the event occur in a small time step $\delta$, divided by the length of the interval. This is formally expressed as [13]

$$h(t) = \lim_{\delta \to 0} \frac{Pr(t < T < t + \delta \mid T > t)}{\delta} \quad (2.15)$$

The Survival function is related to the Hazard Function by

$$h(t) = \frac{f(t)}{S(t)}. \quad (2.16)$$

By defining the *Cumulative Hazard Function* as

$$H(t) = \int_0^t h(u) \, du, \quad (2.17)$$

the relationship to the Survival Function can also be written as

$$S(t) = e^{-H(t)}. \quad (2.18)$$

This relationship comes from the fact that

$$H(t) = \int_0^t h(u) \, du = \int_0^t \frac{f(u)}{S(u)} \, du - \int_0^t \frac{d}{du} \log(S(u)) \, du$$

$$= -\log(S(t)) + \log(S(0)) = -\log(S(t)).$$

### 2.5.2 Right-censoring

A problem with studies during a predetermined time period is that for a group of the subjects in the study the event that is investigated may not occur within the time period. In these cases it is only known that the subjects have survived until the end of the observation period. Even though there
Figure 2.1: An example of right-censoring. The events of subject 3 and 4 occur after the study has ended and hence they are not recorded in the study.

is some bias in the data it is still valuable since it is known how long the subject at least has survived. When a measurement or observation is only partially known like this, the data from the study is called right-censored. It is called right-censored because on a time line the event occurs to the right side of the end of the observation time, see Figure 2.1. [14]

Let $[0, \tau]$ define a time interval where the survival function is investigated. Here 0 can be any time where the investigation starts, i.e. year 1950, and $\tau$ some time later. Let $t_i$ denote the time of renewal of unit $i$, $t_i \in (0, \tau]$. Define the censoring indicator as

$$c_i = \begin{cases} 1 & t_i \leq \tau \\ 0 & t_i > \tau \end{cases} \quad (2.19)$$

The censoring indicator indicates if an event has taken place, e.g. if a turbine has been renewed during the period investigated. This is indicated by $c_i = 1$. Table 2.1 shows an example of how the dataset for lifetimes is structured and how the indicator is used.

Table 2.1 shows an example of how the dataset for lifetimes is structured and how the indicator is used.

The columns plant and unit are self-explaining. YoM stands for year of manufacturing and thus this column indicates which year the unit was started. Renewed indicate which year the unit was taken out of production to be renewed. Dashes in this column indicate that this unit has not been renewed.
before 2016. Years indicate how old the unit was at the time of renewal or is today if this line is censored.

| Plant | Unit | YoM or Renewal | Renewed (Year) | Years | $c_i$
|-------|------|----------------|----------------|-------|-----
| Boden | 1    | 1971           | 1995           | 14    | 1   
| Boden | 1    | 1995           | 2015           | 10    | 1   
| Boden | 1    | 2015           | -              | 1     | 0   
| Boden | 2    | 1972           | 1993           | 21    | 1   
| Boden | 2    | 1993           | 2014           | 21    | 1   
| Boden | 2    | 2014           | -              | 2     | 0   

Table 2.1: Table showing the structure of a dataset. In the $c_i$ column a 1 indicates that an event has taken place during the period of investigation.

### 2.5.3 Non-parametric estimators

Consider the observations $x_1, \ldots, x_n$ from a random sample. If there is no censoring the empirical distribution function is defined as

$$
\hat{F}(t) = \frac{1}{n} \# \{ x_i : x_i \leq x \}
$$

(2.20)

However if the data is censored this has to be taken into account in the function [14]. Let $X = \min(T, C)$ and $c$ be the censoring indicator defined in equation 2.19. The observations are $(x_i, c_i)$ for $i = 1, 2, \ldots, n$ and the likelihood function is

$$
L = \prod_i f(x_i)^{c_i} S(x_i)^{1-c_i} = \prod_i f(x_i)^{c_i} (1 - F(x_i))^{1-c_i}.
$$

Assume there are failure times $0 < t_1 < t_2 < \ldots < t_i < \ldots$. Let $s_{i1}, s_{i2}, \ldots, s_{i\tau_i}$ be the censoring times within the interval $[t_i, t_{i+1}]$ and suppose that there are $d_i$ failures at time $t_i$. The likelihood function then becomes

$$
L = \prod_{F\text{ail}} f(t_i)^{d_i} \prod_i \left( \prod_{k=1}^{\tau_i} (1 - F(s_{ik})) \right)
$$

$$
= \prod_{F\text{ail}} (F(t_i) - F(t_i^-))^{d_i} \prod_i \left( \prod_{k=1}^{\tau_i} (1 - F(s_{ik})) \right)
$$

where the PDF at $t_i$ is written as the difference in the CDF at time $t_i$ and the CDF immediately before it, $f(t_i) = F(t_i) - F(t_i^-)$. 

16
Assuming $F(t_i)$ takes fixed values at the failure time points and using the fact that it is an increasing function, $F(t_i^-)$ and $F(s_{ik})$ are made as small as possible in order to maximize the likelihood. This means taking $F(t_i^-) = F(t_{i-1})$ and $F(s_{ik}) = F(t_i)$. By considering the CDF to be a step function, therefore coming from a discrete distribution with failure times as the actual failure times which occur, will maximizes $L$. The likelihood function is then

$$L = \prod_{Fail} (F(t_i) - F(t_{i-1}))^{d_i} \prod_i (1 - F(t_i))^{\tau_i} \quad (2.21)$$

and it has been shown that the discrete CDF has the maximum likelihood amongst all CDF’s with fixed values $F(t_i)$ at the failure times $t_i$. [14]

Consider the discrete case. Let

$$P(\text{Fail at } t_i|\text{Survived to } t_i^-) = h_i$$

then

$$S(t_i) = 1 - F(t_i) = \prod_{1}^{i} (1 - h_j)$$

and

$$f(t_i) = h_i \prod_{1}^{i-1} (1 - h_j)$$

This finally gives

$$L = \prod_{t_i} h_i^{d_i} (1 - h_i)^{n_i - d_i} \quad (2.22)$$

where $n_i$ is the number at risk at time $t_i$, which indicate how many subjects that have survived until time $t_i$ and are at risk of "dying" within this time step. [14]

### 2.5.3.1 Kaplan-Meier estimator

The Kaplan-Meier estimator, also known as the product limit estimator, is a non-parametric statistic. This estimator uses information from both censored and uncensored data. The Kaplan-Meier estimator comes from the likelihood function in Function 2.22. By taking the log of Function 2.22 the equation

$$\ell = \log(L) = \sum_i d_i \log(h_i) + \sum_i (n_i - d_i) \log(1 - h_i).$$
is obtained [14]. Differentiate with respect to $h_i$ 
\[
\frac{\partial \ell}{\partial h_i} = \frac{d_i}{h_i} - \frac{n_i - d_i}{1 - h_i} = 0 \\
\Rightarrow \hat{h}_i = \frac{d_i}{n_i}
\]

Thus the Kaplan-Meier estimator of the Survival Function is
\[
\hat{S}(t) = \prod_{t_i \leq t} \left( \frac{n_i - d_i}{n_i} \right)
\]
(2.23)

where
\[
n_i = \# \{\text{in risk set at } t_i\}, \\
d_i = \# \{\text{events at } t_i\}.
\]

### 2.5.4 Log-log confidence intervals

Consider the logarithm of the Kaplan-Meier estimator
\[
\log(\hat{S}(t)) = \sum_{t_i \leq t} \log \left( \frac{n_i - d_i}{n_i} \right) = \sum_{t_i \leq t} \log \left( 1 - \hat{h}_i \right)
\]

Assume that the observations in the risk set at time $t_i$ are independent Bernoulli observations with constant probability, then $1 - \hat{h}$ is an estimator of this probability and an estimator of its variance is
\[
\frac{\hat{h}_i(1 - \hat{h}_i)}{n_i} \approx \frac{1}{\mu_X^2} \sigma_X^2 
\]
(2.24)

It can be shown that the log of a variable $X$ is approximately
\[
\text{Var} \left[ \log(X) \right] \approx \frac{1}{\mu_X^2} \sigma_X^2 
\]
(2.25)

where $\mu$ and $\sigma$ are the mean and variance of $X$, respectively. Hence an estimator of the variance is obtained if $\mu$ and $\sigma$ are replace by their estimators,
\[
\text{Var} \left[ \log(1 - \hat{h}_i) \right] \approx \frac{1}{(1 - \hat{h}_i)^2} \frac{\hat{h}_i(1 - \hat{h}_i)}{n_i} = \frac{1}{1 - \hat{h}_i} \frac{\hat{h}_i}{n_i} = \frac{1}{1 - \frac{d_i}{n_i}} \frac{\hat{h}_i}{n_i} \\
= \frac{d_i}{n_i(n_i - d_i)}
\]
Assuming that the observation at each time is independent, the estimator of the variance of the log of the survival function is

$$\hat{\text{Var}} \left[ \log(\hat{S}(t)) \right] = \sum_{t_i \leq t} \hat{\text{Var}} \left[ \log(1 - \hat{h}_i) \right] = \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$  \hspace{1cm} (2.26)

It can be shown that [14]

$$\text{Var}(\exp(X)) \approx (\exp(\mu_X))^2 \sigma_X^2.$$  \hspace{1cm} (2.27)

Let $X$ denote $\log \left[ \hat{S}(t) \right]$, then $\sigma_X^2$ represents the variance estimator in 2.26. Further approximate $\mu_X$ by $\log \left[ \hat{S}(t) \right]$. The expression 2.27 will then be

$$\hat{\text{Var}}(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$  \hspace{1cm} (2.28)

Let $L(t) = \log(-\log(S(t)))$. Start by forming a 95% confidence interval for $L(t)$ based on $\hat{L}(t)$ yielding $[\hat{L}(t) - A, \hat{L}(t) - A]$. Since $S(t) = \exp(-\exp(L(t)))$, the confidence bounds for the 95% confidence interval on $S(t)$ are

$$\left[ \exp\left(-\exp(\hat{L}(t) + A)\right), \exp\left(-\exp(\hat{L}(t) - A)\right) \right].$$

By substituting $\hat{L}(t) = \log(-\log(\hat{S}(t)))$ back into the above bounds, the confidence bound is obtained as

$$\left[ \hat{S}(t)^{\exp(A)}, \hat{S}(t)^{\exp(A)} \right],$$  \hspace{1cm} (2.29)

where $A = 1.96\sqrt{\hat{\text{Var}}(\hat{S}(t))}$ from Equation 2.28. [14]

### 2.5.5 Log-rank test

To test the hazard rates between groups the log-rank test can be used. The hypotheses that are tested are if the hazard rates for $K$ ($K \geq 2$) curves are the same against the alternative that there is at least one that is different for some $t \leq \tau$,

- $H_0$: $h_1(t) = h_2(t) = \ldots = h_K(t), \forall t \leq \tau$
- $H_a$: at least one pair $h_i(t) \neq h_j(t)$ for some $t \leq \tau$.

Here $\tau$ is the largest time at which all of the groups have at least one subject at risk.

Let
$t_1 < t_2 < \ldots < t_D$ be times where events are observed,

- $d_{ik}$ be the number of observed events from group $k$ at time $t_i$,
- $Y_{ik}$ be the number of subjects in group $k$ that are at risk at time $t_i$,

\[
\begin{align*}
    d_i &= \sum_{j=1}^{n} d_{ij}, \\
    Y_i &= \sum_{j=1}^{n} Y_{ij},
\end{align*}
\]

and $W(t_i)$ be the weight of the observations at time $t_i$.

To test the hypothesis a vector $Z$ is computed, where

\[
Z_k = \sum_{i=1}^{D} W(t_i) \left[ d_{ik} - Y_{ik} \frac{d_i}{Y_i} \right],
\]

and the covariance vector $\hat{\Sigma}$, where the variance of $Z_k$ is given by,

\[
\hat{\sigma}_{jj} = \sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i} \left( 1 - \frac{Y_{ij}}{Y_i} \right) \left( \frac{Y_i - d_i}{Y_i - 1} \right) d_i,
\]

and the covariance of $Z_j$ and $Z_g$, $i \neq g$,

\[
\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i} \frac{Y_{ig}}{Y_i} \left( \frac{Y_i - d_i}{Y_i - 1} \right) d_i. \tag{2.30}
\]

The test statistic is then

\[
\chi^2(Z_1, Z_2, \ldots, Z_{K-1}) \Sigma^{-1} (Z_1, Z_2, \ldots, Z_{K-1}) \tag{2.31}
\]

where this statistic has a chi-squared distribution if the null hypothesis is true for large samples with $K - 1$ degrees of freedom. An $\alpha$ level test of $H_0$ rejects when $\chi^2$ is larger than the $\alpha$th upper percentage point of a chi-squared random variable with $K - 1$ degrees of freedom. [13]

The log-rank test is given when the weight function $W(t) = 1$.

2.6 Cox Proportional Hazards model

2.6.1 Regression analysis

Let $Y_i \in \mathbb{R}^1$ and $X_i \in \mathbb{R}^p$. Given a dataset $\{Y_i, X_i\}_{i=1}^{n}$, regression analysis is a statistical method for relating the dependent variable $Y$, often known as *response*, with the independent variables $X$, often knowns as *predictors*. 

20
The regression model also includes unknown parameters $\beta$, that represent how each predictor influence the response.

$$ Y \approx f(\mathbf{X}, \beta) \quad (2.32) $$

To carry out regression analysis an assumption about the form of the function $f$ must be specified. If there is knowledge about the relationship between $Y$ and $\mathbf{X}$ this form is used, otherwise a flexible or simple form for $f$ can be chosen.

Consider $N$ pairs of data points $(\mathbf{X}_i, Y_i)$ and a vector of unknown parameters $\beta$ of length $p$. If $p < N$ it is possible to estimate a unique value for $\beta$ that best fits the data.

### 2.6.2 The model

Cox Proportional Hazards model (also known as Cox model, which will be used interchangeably) is a semi-parametric regression model for the hazard function. To use the model a key assumption has to be fulfilled;

- *Proportional hazards* - The survival curves of two groups must have hazard functions that are proportional over time.

The Cox model is based on a regression model for the hazard function were the hazard is written as a product of two functions;

$$ h(t, x, \beta) = h_0(t)r(x, \beta). \quad (2.33) $$

Here $h_0$ characterizes how the hazard function changes as a function of the survival time, and the function $r(x, \beta)$ characterizes how the hazard function changes as a function of subject covariates. There is a requirement that the functions are chosen such that $h(t, x, \beta) > 0$. When $r(x = 0, \beta) = 1$, $h_0$ is referred to as the baseline hazard function. [14]

Consider the ratio of two hazard function for two subjects with covariates values denoted $x_0$ and $x_1$,

$$ \text{HR}(t, x_0, x_1) = \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)} = \frac{h_0r(x_1, \beta)}{h_0r(x_0, \beta)} = \frac{r(x_1, \beta)}{r(x_0, \beta)}. \quad (2.34) $$

Hence, the hazard ratio (HR) only depends on the function $r(x, \beta)$. The first to propose the model in Equation 2.34 was Cox in 1972 [15]. Cox used $r(x, \beta) = \exp(x\beta)$ and thus got the parametrized hazard function

$$ h(t, x, \beta) = h_0(t)\exp(x\beta) \quad (2.35) $$

and the hazard ratio

$$ \text{HR}(t, x_0, x_1) = \exp(\beta(x_1 - x_0)). \quad (2.36) $$

21
From 2.18 it is known that the survival function is
\[ S(t, x, \beta) = \exp(-H(t, x, \beta)). \] (2.37)

Assume that the survival time is absolutely continuous, this makes it possible to express the cumulative hazard function as
\[ H(t, x, \beta) = \int_0^t h(u, x, \beta) \, du = r(x, \beta) \int_0^t h_0(u) \, du = r(x, \beta)H_0(t) \] (2.38)

The survival function for the general semi-parametric hazard function is then
\[ S(t, x, \beta) = \exp(-r(x, \beta)H_0(t)) \] (2.39)

which can be written as
\[ S(t, x, \beta) = [S_0(t)]^{r(x, \beta)} \] (2.40)

where \( S_0(t) = \exp(-H_0(t)) \) is the baseline survival function [14]. From this the survival function under the Cox model is
\[ S(t, x, \beta) = [S_0(t)]^{\exp(x\beta)}. \] (2.41)

### 2.6.3 Fitting the proportional hazards regression model

Assume there are \( N \) independent observations, each containing a triplet \((t_i, c_i, x_i), i = 1, 2, \ldots, N\) of information. Firstly the triplet give information about the length of time a subject was observed. Secondly if the observation was right censored or survival time. The last information the triplet holds is a single covariate whose value is determined at the time observation begins and remains the same throughout the observation of the subject. The covariate could in the case of hydropower plants be the type of turbine or the effect of the generator.

Use maximum likelihood to estimate the parameters of the proportional hazards model. The first step is to create a specific likelihood function to maximize. Consider the contribution of the triplets \((t, 0, x)\) and \((t, 1, x)\) separately. In the case of \((t, 0, x)\) the survival time was at least \( t \). Hence the contribution from this triplet is the probability that a subject with covariate \( x \) survives at least \( t \) time units, a probability that is given by \( S(t, \beta, x) \). For the triplet \((t, 1, x)\), it is known that the survival time was exactly \( t \), and thus the contribution is the probability that a subject with covariate \( x \) has the event of interest at \( t \) time units.

Further assume that the observations are independent. The likelihood function is then obtained by multiplying the respective contribution from each
triplet over the entire sample. To simplify the calculation the log-likelihood is used and since this function is monotone it is maximized at the same \( \beta \) as the likelihood function. The log-likelihood function is thus

\[
\ell(\beta) = \sum_{i=1}^{N} \left\{ c_i \log \left[ f(t_i, \beta, x_i) \right] + (1 - c_i) \log \left[ S(t_i, \beta, x_i) \right] \right\}.
\] (2.42)

By using the relations in Equation 2.16 and 2.41 the log-likelihood can be rewritten as

\[
\ell(\beta) = \sum_{i=1}^{N} \left\{ c_i \log \left[ h_0(t_i) \right] + c_i x_i \beta + \exp(x_i \beta) \log \left[ S_0(t_i) \right] \right\}.
\] (2.43)

Kalbfleisch and Prentice [16] discussed that the log-likelihood function 2.43 is not possible to use. Cox [15] proposed using a "partial likelihood function" that only depends on the parameter of interest, with the argument that this would have the same distributional properties as the full maximum likelihood estimators. Another problem with the likelihood function 2.43 is that it is not adapted to situations with tied survival times. Efron [17] suggested an approximation that would deal with both of these problems.

Define \( x_{i+} = \sum_{j \in D(t_i)} x_j \), where \( D(t_i) \) represents the set of subjects with survival times equal to \( t_i \). Furthermore assume there are \( m \) distinct ordered survival times, \( t_1 < \ldots < t_m \), and that a number of \( d_i \) events happens at \( t_i \), where \( i = 1, 2, \ldots, m \). Define a set of labels of failing observations at \( t_i \) as \( D(t_i) = \{i_1, \ldots, i_{d_i}\} \). Finally let \( R(t_i) \) denote the set of all subjects of risk at time \( t_i \). This leads to the approximation

\[
L(\beta) = \prod_{i=1}^{m} \frac{\exp(x_{i+} \beta)}{\prod_{k=1}^{d_i} \left[ \sum_{j \in R(t_i)} \exp(x_j \beta) - k-1 \sum_{j \in D(t_i)} \exp(x_j \beta) \right]}.
\] (2.44)

of the likelihood function. [14]

The next step is to find the parameters by maximizing the log-likelihood function in 2.44, but due to cumbersome calculations this will not be presented in this thesis.

### 2.6.4 Verification of the proportional hazards assumption

Since the Cox model assumes proportional hazards it is of great importance to verify that this assumption is fulfilled. One way of doing this is by looking at log cumulative plots.
Let \( \exp(\beta) \) be the proportional hazards constant. Under the proportional hazards assumption the survival function is written as

\[
S_1(t) = [S_0(t)]^{\exp(\beta)}.
\]

Taking log of both sides give

\[
\log (S_1(t)) = \exp(\beta) \log (S_0(t)).
\]

From Equation 2.14 it can be concluded that \( 0 < S(t) < 1 \) must hold. Thus their logarithms are negative and before continuing with the next step they are negated. Take the log of both sides again,

\[
\log \left( -\log (S_1(t)) \right) = \beta + \log \left( -\log (S_0(t)) \right).
\] (2.45)

The function \( \log \left( -\log (g(u)) \right) \) is called the complementary log-log transformation and has the effect of changing the range \((0, 1)\) for \( u \) to \((-\infty, \infty)\) for \( g(u) \). If the proportional hazards assumption is fulfilled a plot of \( g[S_1(t)] \) and \( g[S_0(t)] \) will yield two parallel curves, separated by \( \beta \). [12]

### 2.6.5 Model validation

After the parameters have been calculated and the proportional hazards assumption has been checked it is time to make an assessment of the significance of the model and a confidence interval. There are several test to assess the significance of the parameters; the **partial likelihood ratio test**, the **Wald test** and the **score test**.

Let \( G \) denote the partial likelihood test. It is calculated as two times the difference between the log partial likelihood of the model containing the covariate and the log partial likelihood for the model not containing the covariate,

\[
G = 2 \{ \ell(\hat{\beta}) - \ell(0) \}.
\] (2.46)

Under the null hypothesis that the coefficient is equal to zero, this statistic will follow a chi-square distribution with 1 degree of freedom, \( \chi^2_1 \), and thus can be used to obtain \( p \)-values to test the significance of the coefficient [14].

The Wald test is a parametric statistical test. It tests the true value of the parameter based on the sample estimate. In the Wald test the maximum likelihood estimator of the parameter is compared with the proposed value. The difference between the estimator and the proposed value is assumed to be approximately standard normally distributed. The equation of the Wald statistic is

\[
z = \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})},
\] (2.47)
where $\beta_0 = 0$ in [14] and $\overline{SE}(\hat{\beta})$ is the estimator of the standard error.

The last test is the Score test, also known as the *Lagrange multiplier test*. The Score test tests the hypothesis that the parameter of interest is equal to some particular value,

$$H_0 : \beta = \beta_0$$
$$H_\alpha : \beta \neq \beta_0.$$ 

The statistic to test this hypothesis is

$$S(\beta_0) = \frac{U(\beta_0)^2}{I(\beta_0)},$$ (2.48)

where $S(\beta_0) \rightarrow \chi_1^2$ if the null hypothesis is true. The terms $U(\beta)$ and $I(\beta)$ (expected Fisher information) are

$$U(\beta) = \frac{\partial}{\partial \beta} \ell(\beta)$$
$$I(\beta) = -E \left[ \frac{\partial^2}{\partial \beta^2} \ell(X; \beta) | \beta \right],$$ (2.49)

where $\ell(\beta)$ is the log of the likelihood function 2.44 and $E[X|\beta]$ denotes the expected value of $X$ conditional on $\beta$ [14].

2.7 Aalen’s additive model

Aalen’s additive model is a non-parametric regression model for the hazard function. Unlike the Cox model, the values of the regression coefficients in Aalen’s model are allowed to vary over time.

Consider a vector of $p + 1$ fixed covariates, $x^T = (1, x_1, x_2, \ldots, x_p)$ at some time $t$. The Aalen model then has the hazard function

$$h(t, x, \beta(t)) = \beta_0(t) + \beta_1(t)x_1 + \beta_2(t)x_2 + \ldots + \beta_p x_p.$$ (2.50)

Here the coefficients are time-dependent and provide a change at time $t$, from the baseline hazard rate, $\beta_0(t)$, corresponding to a one-unit change in the respective covariate.

From Equation 2.50 the cumulative hazards function can be calculated by the relation in Equation 2.17,

$$H(t, x, \beta(t)) = \sum_{k=0}^{p} x_k \int_{0}^{t} \beta_k(u) \, du = \sum_{k=0}^{p} x_k B_k(t),$$ (2.51)
where $x_0 = 1$. The term $B_k(t)$ is called the cumulative regression coefficient (CRC) for the $k$:th covariate. From this it follows that $B_0(t)$ is the baseline cumulative hazard function. Aalen noted that it was easier to calculate the CRCs than the regression coefficients [18]. This thesis will use the estimates of the CRCs.

Assume the triplet $\{t_i, x_i, c_i\}$ where $i = 1, \ldots, n$. Each triplet is a unique (no ties) and independent observation of time, with $p$ fixed covariates and a right-censoring indicator variable. Aalen’s estimator of the CRCs is a least-squares-like estimator and is most easily presented using matrices and vectors.

Let $X_j$ denote an $n$ by $p+1$ matrix, where the $i$:th row contains the data for the $i$:th subject, $x_i^T$, if the $i$:th subject is in the risk set at time $t_j$, otherwise all the values of the $i$:th row are zeros.

Let $y_j$ denote a 1 by $n$ vector, where the $j$:th element is 1 if the $j$:th subject’s observed time, $t_j$, is a survival time ($c_j = 1$); otherwise all the values in the vector are zero. The estimator of the CRC at time $t$ is

$$\hat{B}(t) = \sum_{t_j \leq t} \left(X_j^T X_j\right)^{-1} X_j^T y_j.$$  \hspace{1cm} (2.52)

This value of the estimator will only change at observed survival times. The increment in the estimator is computed only when the matrix $\left(X_j^T X_j\right)$ is invertible. When there are fewer than $p+1$ subjects in the risk set the matrix is singular, and therefore cannot be inverted. There is also a possibility of a singular matrix if the model contains a single dichotomous covariate and all subjects who remain at risk have the same value for the covariate. [14]

It follows from Equations 2.51 and 2.52 that the estimator of the cumulative hazard function for the $i$:th subject at time $t$ is

$$\hat{H} \left(t, x_i, \hat{B}(t)\right) = \sum_{k=0}^{p} x_{ik} \hat{B}_k(t)$$ \hspace{1cm} (2.53)

and an estimator of the covariate-adjusted survivor function is

$$\hat{S} \left(t, x_i, \hat{B}(t)\right) = \exp \left[-\hat{H} \left(t, x_i, \hat{B}(t)\right)\right].$$ \hspace{1cm} (2.54)

Aalen noted that it is possible for an estimate of the cumulative hazard Equation 2.53 to be negative and yield a value for Equation 2.54 greater than 1.0. One of the benefits of Aalen’s additive model is to provide graphical covariate-adjusted evidence of the effect of a covariate over time, rather than to provide an additive covariate-adjusted survivor function. [14]
The graphical presentation is most often a plot of $\hat{B}_k(t)$ versus $t$, along with the upper and lower endpoints of a pointwise confidence interval. For a 95% interval Equation 2.55 is used. [14]

$$\hat{B}_k(t) \pm 1.96 \hat{SE}\left[\hat{B}_k(t)\right].$$

(2.55)

$\hat{SE}\left[\hat{B}_k(t)\right]$ is the estimator of the standard error of $\hat{B}_k(t)$, obtained as the square root of the $k$th diagonal element of the variance estimator

$$\hat{\text{Var}} = \left[\hat{B}(t)\right] \sum_{t_j \leq t} \left(X_j^T X_j\right)^{-1} X_j^T I_j X_j \left(X_j^T X_j\right)^{-1}.$$ 

(2.56)
Chapter 3

Analysis

3.1 Condition of the plants

Because of the technical complexity of the units, one of the biggest challenges for the asset managers at Vattenfall Hydro is the assessment of the condition of the power plants. The condition of the power plants is evaluated together with maintenance staff and technical specialists. The assessment of the condition is based on maintenance work during the year, sensor data, failures, etc. The condition is then translated to "good" or "bad" and then used as input data to HYDRA.

Every risk object of the power plant has a list of adverse events, $E_l$, i.e. fire in the generator or bearing failure of turbine, where $l$ denote the type of event. The condition of the parts influence the probabilities of occurrence of the adverse events. There is a hypothesis that there should be a relationship between the condition and the number of occurrences of adverse events. A part $k$ with condition, $S_k$, "bad" should have more occurrences of adverse events, since the probability of failure are higher than for parts with condition "good".

The $k$ conditions are assumed to have equal weight in the total condition of each risk object. Let the $j$:th weight be denoted by $w_j$ and the total number of conditions to be considered are $n$, then the weight of the conditions are

$$w_1 = w_2 = \ldots = w_n = \frac{1}{n}.$$ (3.1)

Thus the condition of each risk object can be assumed to be the mean condition of the $k$ parts,

$$X_i = \sum_{k=1}^{n} w_k S_k = \frac{1}{n} \sum_{k=1}^{n} S_k.$$
Since $S_k \in \{0, 1\}$, $X_k$ will take values in an interval from 0 to 1 with $n - 1$ equally spaced values. From the assumption of the equal weights there will be $n + 1$ groups. These groups can then be used to test for difference between the distribution of some measure in each group.

To test if there are differences between distributions the Kruskal-Wallis test can be used. This test will give an answer to the hypothesis that the distributions of the groups are the same. The alternative hypothesis is that there is some difference between the groups.

If the null hypothesis holds the conclusion can be drawn that there are no differences between the distributions of the measure in each group. This would imply that if the assessments of the conditions are correct there are no differences between the mean of the measure in the groups. However if it is assumed that the occurrences of the measure are correct this instead imply that the assessments of the conditions are incorrect. If instead the null hypothesis is rejected the conclusion can be drawn that there is some difference between the distributions of the measure in each group.

In a case where the null hypothesis is rejected the Spearman rho test can be used to test if the data is correlated in some way and the magnitude of the correlation. If there is positive correlation in the data this imply that for higher values of $x$ there will be higher values of $y$. If the correlation instead is negative this imply that for higher values of $x$ there will be lower values of $y$.

The distributions that will be tested are the availability of the units and the number of hours with unplanned downtime for the units. The null hypotheses in the Kruskal-Wallis test will be that the distribution are the same in each group. This is tested against the alternative hypothesis that $H_0$ is not true.

The availability is the fraction of the number of available hours, $D$, by the total number of hours in a year, $H$ (8760 in a normal year and 8784 in a leap year).

\[ Q = \frac{H - D}{H} = 1 - \frac{D}{H}. \]  \hfill (3.2)

Because the plants need maintenance and upgrades, Vattenfall Hydro distinguishes between planned and unplanned downtime. Planned downtime implies that the downtime of the power plant is planned before the start of the year. Unplanned downtime is caused by one of the following events:

1. There is maintenance work that is planned after the start of the year.
2. The downtime of a planned or unplanned downtime has been prolonged.
3. There is a disruption because of a production-related issue.
4. There is a problem with the network which the electricity is distributed through.

The measure of interest is thus unplanned availability, which will be denoted $Q_{Unplanned}$. The definition of the unplanned availability is

$$Q_{Unplanned} = 1 - \frac{D_{Unplanned}}{H},$$

(3.3)

where $D_{Unplanned}$ denote the number of hours with unplanned downtime.

The other dataset is the number of hours with downtime of a unit due to production-related failures. This dataset can be seen as a subset of the unplanned availability since the number of hours of downtime due to production-related failures is used to calculate the availability.

An assumption will be made that the condition of a unit is the mean condition of the generator and turbine,

$$Y_i = \frac{1}{2} \left( \frac{1}{n} \sum_{k=1}^{n} S_{Generator}^k + \frac{1}{n} \sum_{k=1}^{n} S_{Turbine}^k \right)$$

$$= \frac{1}{2n} \sum_{k=1}^{n} \left( S_{Generator}^k + S_{Turbine}^k \right).$$

3.2 Adverse events

If HYDRA is considered to be ineffective one of the problems could be that the wrong set of adverse events was chosen when the model was created. Therefore it is of great importance to investigate if the adverse events really occur. This can been done in a general fashion by searching for keywords in two different datasets.

The search for keywords has been split into two different cases. Case 1 is general, where the search is just for the specific part that is related to the hazard. In case 2 the search also includes the keyword "Hazard" or "Fire", depending on the adverse event that is searched for.

From these searches information about frequencies can be estimated and it will also be possible to see if the adverse events do occur. The frequencies are calculated as the number of reported events divided by the product of the number of years where there is available data and the number of units,

$$\lambda = \frac{\text{Reported events}}{\text{Years} \cdot \text{Units}}.$$ 

(3.4)
3.3 Probabilities

From the input data of conditions of the power plants together with plant specific properties, i.e. the effect of the generator, HYDRA calculates the risk for each power plant. The risk is the sum of the risks of a predetermined list of adverse events. These adverse events are always the same, and were selected during the process when HYDRA was created.

The risk, Equation 2.1, of an adverse event is the probability of the event to occur multiplied with the cost of restoring the plant back to working condition after the event. The cost, \( c_i \), of restoring the plant after the adverse event, \( E_i \), can be assumed to be deterministic and fixed. Thus the only part of the equation that is variable is the probability \( p_i \) of an adverse event. Figure 3.1 shows a relational chart of what influence the probabilities of adverse events.

![Relational Chart of Influence on Probabilities](image)

Figure 3.1: A relational chart of what influence the probabilities, where the dashed line indicates an indirect relationship.

The probabilities that are used in HYDRA today were estimated by technical specialists during the same process as when the adverse events were chosen. Hence, they are estimated from gut feeling rather than failure frequency. Therefore it is hard to say exactly how well the probabilities are estimated.

As mentioned earlier there have not been any proper recordings of the occurrences of adverse events. Due to this fact it is not possible to estimate the frequencies of adverse event in a satisfactory manner (in Section 3.2 an attempt is made to estimate the frequencies from the dataset of failures in the power plants). There is data available with information about age and renewals of the generators and turbines. However there is a problem with this data, it does not tell if the renewals were the results of an incident or just according to the renewal plan. Therefore an assumption has to be made
that all renewals were the results of hazards.

If reliability analysis is used it is possible to calculate a survival function $S(t)$, Equation 2.14, of the generators and turbines. By using the conditional probability

$$P(t < T < t + 1 | T > t),$$

(3.5)

the probability of a hazard during the coming year can be calculated. If Equation 3.5 is translated into text it is the probability of a hazard during year $(t, t + 1)$ given that the subject has survived until $t$.

Let $A$ be the event $\{ t < T < t + 1 \}$ and $B = \{ T > t \}$. Since $A \subset B$, the intersection of the two sets is $P(A \cap B) = P(A)$. Therefore the conditional probability is

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{P(t < T < t + 1)}{P(T > t)}.$$  

(3.6)

Equation 3.6 can then be used to calculate the probabilities of a hazard for all years the survival function is valid for. These probabilities can then be compared with the probabilities in HYDRA.

The probabilities in HYDRA are divided into sets for each risk object. To compare these probabilities with the calculated ones in Equation 3.6 the union of probabilities of adverse events $E_k$ can be used. If the probabilities of adverse events are considered to be disjoint this will be the sum of the probabilities,

$$P(E_1 \cup E_2 \cup \ldots \cup E_n) = \sum_k P(E_k).$$

The Kaplan-Meier estimator will be used to calculate the survival function for the generators and turbines. To test how the survival depends on covariates two regression models will be used; Cox proportional hazards model and Aalen’s additive regression model. From the regressions it will be possible to see if the model of survival can be improved by considering the covariates. There are two covariates for both the generators and turbines and all of them are dichotomous, see Table 3.1. The probabilities from Aalen’s and Cox model will also be compared with the probabilities used in HYDRA today.

<table>
<thead>
<tr>
<th>Year of manufacturing</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>&gt; 1960</td>
</tr>
<tr>
<td>1</td>
<td>\leq 1960</td>
</tr>
</tbody>
</table>

Table 3.1: The covariates used in the regressions.
3.4 Risk forecasts

Even though HYDRA was not built to predict upcoming events it is of interest to see if risk assessments from HYDRA could indicate this. In the section of risk forecasts the models historical predictability of risks will be treated. How well have the model predicted risks during the ten years it has been used?

To answer this question it is important to understand what the risks are. As stated earlier in Equation 2.1 risks are products of probabilities and consequences of adverse events. Since the data regarding adverse events is imperfect other data has to be considered. One type of available data is information about failure events that have led to insurance claims.

The insurance claims have information about damaged objects. Using this information together with the risk assessments from HYDRA it should be possible to answer the question about the models risk predicting for certain cases. If the model can predict these kind of events the condition of the plants should be worse and the risk higher.

3.5 Consequences and trends

One of the most important things to investigate regarding HYDRA is if the information Vattenfall Hydro gets from the model gives them valuable knowledge that help them with planning of maintenance and renewals. Assuming that the model gives reliable information, it is of interest to investigate if there are consequences from the plans made given the model results.

In short term there are not many options to make big improvements of the condition of a unit. The big improvements comes from CAPEX. The alternative is to use OPEX, but it is not known how big the impact is on the condition from these actions. Hence it is of interest to see if OPEX is enough to improve the condition of each unit.

Since the condition of the units is estimated once a year it has to be assumed that the condition is fixed and cannot change during the year. Take the difference in condition of each unit year $i$ and year $i - 1$, $S_{i}^{\text{Unit}} - S_{i-1}^{\text{Unit}}$. This gives the change in condition and can be compared with the OPEX during year $i - 1$.

Here the Kruskal-Wallis test can be used with the hypothesis

$$H_0: OPEX_1 \overset{d}{=} OPEX_2 \overset{d}{=} \ldots \overset{d}{=} OPEX_{n+1}, \quad (3.7)$$
where \( i + 1 \) is the number of groups with difference in condition. In the case where \( H_0 \) can be rejected the Spearman test can be used to test the magnitude of the correlation, and if it is positive or negative.

By reasoning it is natural to assume that if OPEX can improve the condition it should be possible to see a positive correlation between the difference in condition and OPEX, where a larger input of OPEX year \( i - 1 \) leads to an improved condition year \( i \).

One way of further investigating consequences of actions taken is to investigate trends during the years since HYDRA was introduced. These trends will not tell if HYDRA is performing as it should, but they should give indications if the information that the model provides is used in the correct way. Some metrics that are of interest are the trends of spill, lost margin, availability and the number of hours with interruptions due to production-related failures.
Chapter 4

Results

In this chapter the results of the thesis are presented. The chapter is divided into five sections, one for each question to be answered. Firstly considered is the condition of the plants, and then the effects from actions taken. After that the reliability of the probabilities is presented followed by the results in the investigation of the adverse events. Finally the investigation of risk forecasts is presented. If not stated, the significance level is $\alpha = 5\%$ in a statistical test.

4.1 Condition of the plants

As mentioned earlier the model depends on the estimations of the conditions of the parts, which is translated to a condition "good" or "bad". To learn how well Vattenfall Hydro is at estimating the condition of the parts the mean condition of the risk objects will be compared with two datasets:

1. Availability of each unit
2. Hours of downtime due to production-related interruptions

4.1.1 Condition against the availability

The first dataset to investigate is the availability of the units. In Figure 4.1 is a scatter plot of the mean condition of units against the availability. As in the previous case it is difficult to draw any conclusions from the plot.

In Table 4.1 are the results from the Kruskal-Wallis test. From the results the conclusion can be drawn that there are no difference between the dis-
Figure 4.1: The mean condition per unit against the availability in percentage.

tributions. Hence the means are equal and it is not possible to relate the availability to the condition of the units.

<table>
<thead>
<tr>
<th>$Q_{\text{Units Unplanned availability}}^{\text{Units}}$</th>
<th>$F_{KW}$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.4301</td>
<td>0.3494</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Results of the Kruskal-Wallis test for the availability of the groups of conditions.

4.1.2 Condition against hours of downtime due to production-related interruptions

The second dataset to investigate the number of hours of downtime due to production-related interruptions. Figure 4.2a shows a scatter plot of the mean condition of units against the number of hours of production-related interruptions. As previously it is difficult to draw conclusions by investigating the plot.

The results from the Kruskal-Wallis test can be seen in Table 4.2. $F_{KW}$ is slightly larger than the $\chi^2$-value with 14 degrees of freedom and $\alpha = 5\%$, 23.645, which is also confirmed by the $p$-value. Hence the null hypothesis is rejected and there are differences between the distributions. Thus it is not possible ignore the fact that there is a possibility to relate hours with inter-
(a) The mean condition per unit against the number of hours of downtime due to production-related interruptions.

(b) Quantiles for the number of hours of downtime due to production-related interruptions for each group of mean condition of the units.

Figure 4.2: A scatter plot showing the mean condition of the units against the number of production-related interruptions, and a box plot showing the quantiles of hours of downtime due to production-related interruptions for each group of mean condition of the units.

ruptions to the condition of the units. A Spearman test can be performed to investigate if there is a correlation and if it is positive or negative.

<table>
<thead>
<tr>
<th>$D_{\text{Units}}^{\text{Production-related interrupts}}$</th>
<th>$F_{\text{KW}}$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.3947</td>
<td>0.04103</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: The result of the Kruskal-Wallis test of the number of hours of production-related interruptions for the groups of conditions.

<table>
<thead>
<tr>
<th>$D_{\text{Units}}^{\text{Production-related interrupts}}$</th>
<th>Correlation coefficient $\rho$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05591</td>
<td>0.02188</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: The results of the Spearman test of the number of hours of production-related interruptions for the groups of conditions.

In Table 4.3 are the results of the Spearman test. The test shows that the correlation is positive and significant but very small, $\rho = 0.056$. This result
Table 4.4: The keywords used in the different searches. Here the part related to adverse event is the main keyword.

goes against what is expected as the logical thing would be that units with good condition should have fewer hours with interruptions.

4.2 Adverse events

HYDRA is built on a list of adverse events. To see if these events occur, the number of reported events and the frequencies has been calculated. The frequency is represented by $\lambda$. The results are divided into four parts of two datasets and two cases in each dataset. In each search the name of the part related to the adverse event was used as a keyword. The difference between Case 1 and Case 2 is that an extra keyword, hazard, is added in the search of events in Case 2.

4.2.1 Dataset 1

Dataset 1 is the set of all reported events in the power plants during the period 2004 to 2015.

4.2.1.1 Case 1

For case 1 of dataset 1 the results can be seen in Table 4.5 and 4.6. There are records of reported events for each keyword.

4.2.1.2 Case 2

In Tables 4.7 and 4.8 are the results of the searches for keywords in combination with "hazard". These numbers differ significantly from the numbers in Tables 4.5 and 4.6. In the cases where there are 0 reported events the frequency is unspecified but assumed to be lower than if there would have been one reported event.
Table 4.5: The data in this table is from dataset 1, case 1 for the generators. The table presents the number of reported events and the frequency $\lambda$.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Reported events</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stator</td>
<td>268</td>
<td>0.1718</td>
</tr>
<tr>
<td>Rotor</td>
<td>162</td>
<td>0.1038</td>
</tr>
<tr>
<td>Magnetization</td>
<td>287</td>
<td>0.1840</td>
</tr>
<tr>
<td>Bearing</td>
<td>1559</td>
<td>0.9994</td>
</tr>
<tr>
<td>Fire</td>
<td>36</td>
<td>0.0231</td>
</tr>
</tbody>
</table>

Table 4.6: The data in this table is from dataset 1, case 1 for the turbines. The table presents the number of reported events and the frequency $\lambda$.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Reported events</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runner</td>
<td>215</td>
<td>0.1378</td>
</tr>
<tr>
<td>Hub</td>
<td>11</td>
<td>0.0071</td>
</tr>
<tr>
<td>Vane</td>
<td>108</td>
<td>0.0692</td>
</tr>
<tr>
<td>Bearing</td>
<td>1036</td>
<td>0.6814</td>
</tr>
<tr>
<td>Servo</td>
<td>132</td>
<td>0.0846</td>
</tr>
</tbody>
</table>

4.2.2 Dataset 2

Dataset 2 is the set of all reported production-related events in the power plants during the period 2012 to 2015. In Case 2, there is only one occurrence on the search of keywords. Hence the frequency is calculated for one event for each keyword and then the frequency is assumed to be smaller than this calculated number.
<table>
<thead>
<tr>
<th>Keyword + hazard</th>
<th>Reported events</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stator</td>
<td>0</td>
<td>0.000641</td>
</tr>
<tr>
<td>Rotor</td>
<td>0</td>
<td>&lt; 0.000641</td>
</tr>
<tr>
<td>Magnetization</td>
<td>1</td>
<td>0.000641</td>
</tr>
<tr>
<td>Bearing</td>
<td>3</td>
<td>0.001923</td>
</tr>
<tr>
<td>Fire</td>
<td>2</td>
<td>0.001282</td>
</tr>
</tbody>
</table>

Table 4.7: The data in this table is from dataset 1, case 2 for the generators. The table presents the number of reported events and the frequency $\lambda$.

<table>
<thead>
<tr>
<th>Keyword + hazard</th>
<th>Reported events</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runner</td>
<td>1</td>
<td>0.000641</td>
</tr>
<tr>
<td>Hub</td>
<td>1</td>
<td>0.000641</td>
</tr>
<tr>
<td>Vane</td>
<td>0</td>
<td>&lt; 0.000641</td>
</tr>
<tr>
<td>Bearing</td>
<td>1</td>
<td>0.000641</td>
</tr>
<tr>
<td>Servo</td>
<td>2</td>
<td>0.001282</td>
</tr>
</tbody>
</table>

Table 4.8: The data in this table is from dataset 1, case 2 for the turbines. The table presents the number of reported events and the frequency $\lambda$.

### 4.2.2.1 Case 1

This dataset is much smaller than dataset 1, which easily is seen by the number of reported events that are found when search for the same keywords as in dataset 1, see Tables 4.9 and 4.10. However the frequencies are quite large. Since there are about 130 units in Vattenfall’s hydropower plant fleet this would mean that for all keywords, except one, there will be at least one hazard each year. In some cases there would even be more than that.
<table>
<thead>
<tr>
<th>Keyword</th>
<th>Reported events</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stator</td>
<td>17</td>
<td>0.03269</td>
</tr>
<tr>
<td>Rotor</td>
<td>10</td>
<td>0.01923</td>
</tr>
<tr>
<td>Magnetization</td>
<td>24</td>
<td>0.04615</td>
</tr>
<tr>
<td>Bearing</td>
<td>115</td>
<td>0.2212</td>
</tr>
<tr>
<td>Fire</td>
<td>6</td>
<td>0.1154</td>
</tr>
</tbody>
</table>

Table 4.9: Data from dataset 2, case 1 for the generators. The table shows the number of reported events and the frequency $\lambda$.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Reported events</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runner</td>
<td>19</td>
<td>0.03654</td>
</tr>
<tr>
<td>Hub</td>
<td>3</td>
<td>0.005769</td>
</tr>
<tr>
<td>Vane</td>
<td>7</td>
<td>0.01346</td>
</tr>
<tr>
<td>Bearing</td>
<td>115</td>
<td>0.2212</td>
</tr>
<tr>
<td>Servo</td>
<td>8</td>
<td>0.01539</td>
</tr>
</tbody>
</table>

Table 4.10: Data from dataset 2, case 1 for the turbines. The table shows the number of reported events and the frequency $\lambda$.

### 4.2.2.2 Case 2

In the last case only one reported event was found with the search for keywords in combination with "hazard", see Tables 4.11 and 4.12. In the cases where there are 0 reported events the frequency is unspecified but assumed to be lower than if there would have been one reported event.


<table>
<thead>
<tr>
<th>Keyword + hazard</th>
<th>Reported events</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stator</td>
<td>0</td>
<td>$&lt; 0.001923$</td>
</tr>
<tr>
<td>Rotor</td>
<td>0</td>
<td>$&lt; 0.001923$</td>
</tr>
<tr>
<td>Magnetization</td>
<td>0</td>
<td>$&lt; 0.001923$</td>
</tr>
<tr>
<td>Bearing</td>
<td>1</td>
<td>$0.001923$</td>
</tr>
<tr>
<td>Fire</td>
<td>0</td>
<td>$&lt; 0.001923$</td>
</tr>
</tbody>
</table>

Table 4.11: The data in this table is from dataset 2, case 2 for the generators. The table presents the number of reported events and the frequency $\lambda$.

<table>
<thead>
<tr>
<th>Keyword + hazard</th>
<th>Reported events</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runner</td>
<td>0</td>
<td>$&lt; 0.001923$</td>
</tr>
<tr>
<td>Hub</td>
<td>0</td>
<td>$&lt; 0.001923$</td>
</tr>
<tr>
<td>Vane</td>
<td>0</td>
<td>$&lt; 0.001923$</td>
</tr>
<tr>
<td>Bearing</td>
<td>0</td>
<td>$&lt; 0.001923$</td>
</tr>
<tr>
<td>Servo</td>
<td>0</td>
<td>$&lt; 0.001923$</td>
</tr>
</tbody>
</table>

Table 4.12: The data in this table is from dataset 2, case 2 for the turbines. The table presents the number of reported events and the frequency $\lambda$.

4.3 Probabilities

In this section the results of the comparison between the probabilities of the model and probabilities estimated by survival analysis are presented.

4.3.1 Kaplan-Meier estimator

To estimate the survival functions of the datasets the Kaplan-Meier estimator is used. In Figure 4.3 are the estimators of the generator and turbine datasets. These estimates are made without considering any differences between certain groups in the data. In the generator dataset there are 229
records and 101 of these are events, meaning that they are not censored. The corresponding numbers of the turbine dataset are 339 records and 209 events. The median survival time is 49 years for generators and 27 years for turbines.

Table 4.13: Grouping of the turbines.

| Effect \(\leq 55\) MW | Manufactured \(\leq 1960\) | Group 1 |
| Effect \(> 55\) MW | Manufactured \(> 1960\) | Group 2 |

(a) Kaplan-Meier estimator without covariates of the generators.

(b) Kaplan-Meier estimator without covariates of the turbines.

Figure 4.3: Survival functions by Kaplan-Meier estimates of generators and turbine.

It is of interest to test if there are differences between certain groups in the datasets. To group the data there are two covariates for each dataset. In the dataset of the generators there are two columns telling the year of manufacturing and the rated capacity. In the dataset of the turbines there are two columns telling the year of manufacturing and what type of turbine it is. In Table 4.13 and 4.14 are the groups of the generators and turbines respectively.

To test if there are differences between groups, a log-rank test is performed on the hazard functions. In Table 4.15 are the results from the log-rank tests and in Figures 4.4a and 4.4b are plots of the survival functions. The results of the log-rank test indicate that the null hypothesis can be rejected on the
<table>
<thead>
<tr>
<th></th>
<th>Manufactured ≤ 1960</th>
<th>Manufactured &gt; 1960</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaplan</td>
<td>Group 1</td>
<td>Group 3</td>
</tr>
<tr>
<td>Francis</td>
<td>Group 2</td>
<td>Group 4</td>
</tr>
</tbody>
</table>

Table 4.14: Grouping of the turbines.

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generators</td>
<td>16.3</td>
<td>0.00101</td>
</tr>
<tr>
<td>Turbines</td>
<td>56.5</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 4.15: Results from the log-rank test for the generators and the turbines.

$\alpha = 5\%$ level. If the plots are examined more closely there are one group for both the generators and turbines that have too few records of events.

(a) Kaplan-Meier estimator of each group in the generator data.

(b) Kaplan-Meier estimator of each group in the turbine data.

Figure 4.4: Kaplan-Meier estimators for the generators and turbines.

In Tables 4.16 and 4.17 are the records and events of the analyses with covariates presented. The records include censored events and the number of events are the records without the censored events. Here there is one column in each table were there seems to be too few data points. Group 1 for the generators only has 6 events registered and Group 4 for the turbines only has 7 data points in total.
<table>
<thead>
<tr>
<th>Group</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records</td>
<td>61</td>
<td>87</td>
<td>56</td>
<td>25</td>
</tr>
<tr>
<td>Events</td>
<td>6</td>
<td>27</td>
<td>45</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 4.16: The number of records and events in each group of the generators.

<table>
<thead>
<tr>
<th>Group</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records</td>
<td>121</td>
<td>83</td>
<td>48</td>
<td>7</td>
</tr>
<tr>
<td>Events</td>
<td>53</td>
<td>52</td>
<td>46</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 4.17: The number of records in each group of the turbines.

Due to the lack of data points the grouping will only be done on one covariate at time. In Table 4.18 are the results of log-rank test on each grouping. Here it is not possible to reject the null hypothesis of the grouping on the year of manufacturing for the generators. This becomes quite obvious by looking at a plot of the estimates of the survival functions for each group, see Figure 4.5a.

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>YoM (Generators)</td>
<td>0.4</td>
<td>0.554</td>
</tr>
<tr>
<td>Effect (Generators)</td>
<td>12.8</td>
<td>0.001</td>
</tr>
<tr>
<td>YoM (Turbines)</td>
<td>44.3</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Type (Turbines)</td>
<td>27</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 4.18: Results from log-rank tests for the generators and the turbines.

In Figure 4.5b is the plot of the Kaplan-Meier estimate for the grouping on effect. Here the results from the log-rank test can be confirmed graphically.

In Figure 4.6 are the Kaplan-Meier estimates for both groupings of the turbines. As for the effect of the generators, both groupings are relevant in
(a) Kaplan-Meier estimators for the generators where the grouping is made on year of manufacturing.

Figure 4.5: Kaplan-Meier estimates of the generators.

(b) Kaplan-Meier estimators for the generators where the grouping is made on effect of generator.

estimating the survival of the turbines.

(a) Kaplan-Meier estimators for the turbines where the grouping is made on year of manufacturing.

(b) Kaplan-Meier estimators for the turbines where the grouping is made on type of turbine.

Figure 4.6: Kaplan-Meier estimates of the turbines.

4.3.2 Two regression models

To assess if the covariates improve the model of the survival regression analysis can be used. In this section follows two approaches of regression analysis,
Aalen’s additive model and the Cox model.

4.3.2.1 Aalen’s additive model

In Aalen’s additive model the effect from the covariates is allowed to be time-dependent. To assess if a covariate should be included in the model plots of the cumulative regression functions against the time for each covariate can be graphically inspected. If the covariate is time-dependent a plot showing that the cumulative regression function diverges from zero is expected. If instead the covariate is not time-dependent a Brownian motion like behavior around zero is expected.

In Figure 4.7 are two plots of the cumulative regression functions for the generators. Here it can clearly be seen that the covariate Early is not time-dependent, Figure 4.7a. High Effect, however, is time-dependent as Figure 4.7b shows that the cumulative regression function moves away from zero.

If the results of the regression analysis for the covariate Early is inspected it shows that Early is not significant, see Table 4.19. The $p$-value is very big and 0 is included in the confidence interval. The results from the regression analysis with only one covariate can be seen in Table 4.20. Here both the constant and the covariate High Effect are significant. For the turbines the cumulative regression function of the covariates can be seen in Figure 4.8.

From looking at the plots it seems like both covariates are time-dependent. However, if the results of the regression analysis are considered it shows that 0 is in the confidence interval for Type, see Table 4.21. Hence Type should be excluded from the model. The results from the regression analysis of the

![Figure 4.7: Cumulative regression functions for the generators.](image-url)
Table 4.19: Regression results from Aalen’s additive regression with the covariate *Early* for the year of manufacturing of the generators.

<table>
<thead>
<tr>
<th>Coeff.</th>
<th>Std. Error</th>
<th>95% C.I.</th>
<th>Test stat.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>0.037</td>
<td>0.296</td>
<td>[−0.543, 0.617]</td>
<td>−0.136</td>
</tr>
</tbody>
</table>

Table 4.20: Regression results from Aalen’s additive regression with the covariate for the effect of the generators.

<table>
<thead>
<tr>
<th>Coeff.</th>
<th>Std. Error</th>
<th>95% C.I.</th>
<th>Test stat.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.050</td>
<td>0.191</td>
<td>[0.675, 1.425]</td>
<td>6.633</td>
</tr>
<tr>
<td>High Effect</td>
<td>1.529</td>
<td>0.612</td>
<td>[0.329, 2.729]</td>
<td>3.185</td>
</tr>
</tbody>
</table>

(a) Cumulative regression function of the covariate for year of manufacture.  
(b) Cumulative regression function of the covariate for the effect.

Figure 4.8: Cumulative regression functions for the turbines.

<table>
<thead>
<tr>
<th>Coeff.</th>
<th>Std. Error</th>
<th>95% C.I.</th>
<th>Test stat.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>2.500</td>
<td>1.358</td>
<td>[−0.162, 5.161]</td>
<td>2.885</td>
</tr>
</tbody>
</table>

Table 4.21: Regression results from Aalen’s additive regression model with the covariate for the type of turbine.

final model with only one covariate, *Early*, can be seen in 4.22.
<table>
<thead>
<tr>
<th></th>
<th>Coeff.</th>
<th>Std. Error</th>
<th>95% C.I.</th>
<th>Test stat.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>4.488</td>
<td>1.258</td>
<td>[2.021, 6.955]</td>
<td>10.247</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Early</td>
<td>−3.559</td>
<td>1.269</td>
<td>[−6.046, −1.071]</td>
<td>−6.401</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 4.22: Regression results from Aalen’s additive regression model with the covariate for the year of manufacturing of the turbines.

4.3.2.2 The Cox model

To use the Cox model the assumption of proportional hazards have to be fulfilled. Hence this is the first thing to consider. This can be done graphically by looking at log cumulative hazard plots. In Figures 4.9a and 4.9b are the log cumulative plots for the generator data. The crossing lines indicate that the assumption of proportional hazards is not fulfilled. This conclusion can be drawn since proportional hazards would mean that the lines are parallel everywhere. Hence the Cox model cannot be used for this data. If instead

(a) Log cumulative hazard plots of the year of manufacturing for the generator.

(b) Log cumulative hazard plots of the effect for the generator.

Figure 4.9: Log cumulative hazards plots for the generators.

the turbines are considered the results of the log cumulative hazards plot can be seen in Figures 4.10a and 4.10b. Unlike the generators, the lines in these figures do not cross each other. However, the distance between the lines is not constant in any of the figures. In this case it is hard to tell if the Cox model can be applied or not. The assumption will be assumed to be fulfilled and hence the Cox model can be applied to the turbine data. Applying the
Cox model with the covariates for manufacturing year and turbine type gives the results in Table 4.23. Both covariates is significant at $\alpha = 5\%$ level. In Table 4.24 are the results of the test statistics. Applying the Cox model significantly improves the modelling of the survival of the turbines.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>$\exp(\beta)$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>-1.2386</td>
<td>0.2898</td>
<td>0.776</td>
</tr>
<tr>
<td>Type</td>
<td>0.5720</td>
<td>1.7718</td>
<td>0.00188</td>
</tr>
</tbody>
</table>

Table 4.23: The regression coefficients, including the $p$-values, for the Cox model of the turbine data.

<table>
<thead>
<tr>
<th></th>
<th>Likelihood ratio test</th>
<th>Wald test</th>
<th>Score (logrank) test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
<td>56.73</td>
<td>51.68</td>
<td>56.71</td>
</tr>
<tr>
<td>$p$-value</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 4.24: Test statistics of the Cox model of the turbine data.
4.3.3 Comparisons of the probabilities

In this section the models from Section 4.3.1 and 4.3.2 are used to calculate the probability of a hazard during the upcoming year given that there has been no hazards up until today.

In Table 4.25 are the results for the generators. The results from the different models do not differ that much. Two results that may seem strange are the results for Kaplan Meier for age 5 and 40. In the first case, the reason for the probability of 0 is because there were no events in the data before the age of 7. For the other case the reason is due to the fact that if there are many events during the same time step, the probability will be larger of having an incident in these periods.

The probabilities in HYDRA are constant in time, only depending on the condition of the plants and the attributes. If all conditions in the plants are good the probability of an hazard in a generator are \( \sim 0.00003 \). If instead all conditions in the plants are bad the probability of an hazard in a generator \( \sim 0.2 \). The reason for the big difference between probabilities is because the probabilities were not chosen to reflect the real probability of a hazard. Instead they were chosen to highlight changes in the conditions of the plants. The same holds for the probabilities of a hazard in the turbines.

<table>
<thead>
<tr>
<th>Age</th>
<th>KM (( \leq 1960 ))</th>
<th>KM (&gt;1960)</th>
<th>Aalen’s additive model</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>0.00760</td>
<td>0.00966</td>
</tr>
<tr>
<td>10</td>
<td>0.0127</td>
<td>0.00890</td>
<td>0.0117</td>
</tr>
<tr>
<td>20</td>
<td>0.0263</td>
<td>0.0109</td>
<td>0.0290</td>
</tr>
<tr>
<td>40</td>
<td>0.0188</td>
<td>0.0454</td>
<td>0.0657</td>
</tr>
</tbody>
</table>

Table 4.25: Probabilities of having an incident at the age of 5, 10, 20 and 40 years from two different survival models for the generators.

If all conditions are good the probability of an hazard is \( \sim 0.0002 \) and if they are all bad the probability is \( \sim 0.5 \). These probabilities can be compared with the calculated probabilities in Table 4.26.
### Table 4.26: Probabilities of having an incident at the age of 5, 10, 20 and 40 years from three different survival models for the turbines.

<table>
<thead>
<tr>
<th>Age</th>
<th>KM(≤1960)</th>
<th>KM (&gt;1960)</th>
<th>Aalen’s additive model</th>
<th>Cox Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>0.0119</td>
<td>0.0180</td>
<td>0.00789</td>
</tr>
<tr>
<td>10</td>
<td>0.0192</td>
<td>0.0424</td>
<td>0.0190</td>
<td>0.0339</td>
</tr>
<tr>
<td>20</td>
<td>0.0213</td>
<td>0.0735</td>
<td>0.0206</td>
<td>0.0458</td>
</tr>
<tr>
<td>40</td>
<td>0.0833</td>
<td>0.333</td>
<td>0.0714</td>
<td>0.0837</td>
</tr>
</tbody>
</table>

4.4 Risk forecasts

In this section four different events are investigated closer by looking at the output from HYDRA and metrics of the number of events, the availability and the number of hours of outage due to production-related issues.

4.4.1 Boden 2013

The 28th August 2013 unit G1 in Boden had an incident. The incident was a rotor breakdown and resulted in a outage of the unit. In Figure 4.11 the calculated risk from HYDRA is presented. Here it is easy to spot the increase in risk 2012. The large risk was due to condition "bad" on the stator, rotor and bearing. There was also a comment which said to have this unit under supervision because Vattenfall knew there was a problem with it. The risk for 2013 was calculated after the incident. In this case it seems like HYDRA

![Figure 4.11: Yearly risk of Boden G1 from 2004 until 2013.](image)
could forecast the increased risk. The changes in condition of vital parts made the risk increase and the bad condition of the parts eventually lead to an incident. The age of the generator year 2013 was 41 and the generator had an effect of 50 MW. In Table 4.27 is the number of reported events for the generator during the period 2009 until 2013. There is no observations, leading up to the incident, that indicate a future incident. In 2012 there were no registered events for the generator in G1. Thus it would have been difficult foreseeing the hazard by looking at the number of reported events. Presented in the same table is also the availability and hours with interruptions of the same period. There is little evidence in these numbers indicating there is an upcoming larger incident.

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of reported incidents</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>Availability [%]</td>
<td>98.72</td>
<td>96.28</td>
<td>93.84</td>
<td>98.79</td>
<td>64.44</td>
</tr>
<tr>
<td>Hours with interruptions</td>
<td>0</td>
<td>245</td>
<td>366</td>
<td>0</td>
<td>2937</td>
</tr>
</tbody>
</table>

Table 4.27: Number of reported events, availability and hours with interruptions of the generator in Boden unit G1 during the period 2009 until 2013.

4.4.2 Juktan 2011

The next incident happened in the power plant Juktan the 14th December 2011. Unit G1 in Juktan had an incident on the generator. If Figure 4.12 is examined there is nothing in the yearly risk that indicated an upcoming incident. In Table 4.28 are some measurements for unit G1. The number of reported events in more than normal for the period, but not significantly higher. If the availability and hours with interruptions is investigated the numbers for 2011 show that the condition of the units was worse than expected. The unplanned availability was only 59.79 % this year and the hours with interruptions were unusually high. Even though the incident influences these numbers, it should only be to a minor degree since the accident happened in the middle of December.

4.4.3 Olidan 2010

The third incident happened in the power plant Olidan unit G8 the 5th May 2010. This unit first used in production 2009 and hence there is not
Figure 4.12: Yearly risk of Juktan G1 from 2004 until 2013.

<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of reported incidents</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Availability [%]</td>
<td>92.48</td>
<td>87.58</td>
<td>88.17</td>
<td>94.12</td>
<td>59.79</td>
</tr>
<tr>
<td>Hours with interruptions</td>
<td>12</td>
<td>43</td>
<td>980</td>
<td>37</td>
<td>2959</td>
</tr>
</tbody>
</table>

Table 4.28: Number of reported events, availability and hours with interruptions of the generator in Juktan unit G1 during the period 2007 until 2011.

much information before the incident. In Figure 4.13 are the risks from year 2009 to 2013. The condition of the plant was assessed a short time before the incident, with the result of a decreased risk from 2009 to 2010. If the measures of incident, availability and hours with interruptions are inspected there is nothing indicating an upcoming incident.

4.4.4 Torpshammar 2009

The last incident to be analyzed happened at the power plant Torpshammar 23rd December 2009, in unit G2. The incident inflicted damage on a generator fan blade, on rotor top and windings. By inspecting Figure 4.14 the risk during the period was higher than normal. There are also comments in the HYDRA-report telling that the unit is not in good condition. The condition of unit G2 is not reflected in Table 4.30. According to these measures everything seems to have been normal up until the incident.
Table 4.29: Number of reported events, availability and hours with interruptions of the generator in Olidan unit G8 during the period 2009 until 2010.

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of reported incidents</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Availability [%]</td>
<td>100</td>
<td>85.67</td>
</tr>
<tr>
<td>Hours with interruptions</td>
<td>0</td>
<td>1255</td>
</tr>
</tbody>
</table>

4.5 Consequences and trends

If it is assumed that the information from HYDRA is correct it is of interest to see if this information is used in a correct way. Here follows results from a
2005 2006 2007 2008 2009

<table>
<thead>
<tr>
<th>Number of reported incidents</th>
<th>6</th>
<th>7</th>
<th>5</th>
<th>2</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability [%]</td>
<td>99.35</td>
<td>99.68</td>
<td>99.54</td>
<td>99.53</td>
<td>96.41</td>
</tr>
<tr>
<td>Hours with interruptions</td>
<td>1</td>
<td>5</td>
<td>24</td>
<td>14</td>
<td>237</td>
</tr>
</tbody>
</table>

Table 4.30: Number of reported events, availability and hours with interruptions of the generator in Torpshammar unit G2 during the period 2005 until 2009.

comparison between OPEX against the change in condition from year \( i - 1 \) to year \( i \) and some aggregate trends during the time HYDRA has been used.

### 4.5.1 OPEX against difference in condition

In this section are the results from the test to investigate whether it is possible to improve the condition of the units in the hydropower plants by OPEX. This test was performed by looking at changes in condition from year \( i - 1 \) to year \( i \) against OPEX from year \( i - 1 \). In Figure 4.15a is a scatter plot of the difference in condition against the OPEX. By only looking at the figure it is hard to tell if there is a positive correlation. To investigate this the Kruskal-Wallis test was used. The \( \chi^2 \)-value for 38 degrees of freedom with \( \alpha = 5\% \) is 53.38354 which is smaller than \( F_{KW} \), Table 4.31. This is also confirmed by the \( p \)-value. Thus the null hypothesis is rejected and it is possible to conclude that there are differences between the distributions of the groups.

<table>
<thead>
<tr>
<th>OPEX</th>
<th>( F_{KW} )</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>82.4396</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 4.31: The result of the Kruskal-Wallis test for OPEX in each group of difference in condition.

The results of the Spearman test can be seen in Table 4.32. From these results it is possible to conclude that there is a small but positive correlation between OPEX and the difference in condition. Hence there is a small possibility to improve the condition by OPEX.

58
Change in mean condition of all units in each plant against OPEX from preceding year

(a) OPEX year \(i - 1\) against the difference in condition from year \(i - 1\) to \(i\).

(b) The quantiles of OPEX for each group of condition of the units.

Figure 4.15: A scatter plot and a box plot showing the difference in condition between years \(i - 1\) and \(i\) against OPEX.

<table>
<thead>
<tr>
<th>Correlation coefficient (\rho)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPEX</td>
<td>0.1531</td>
</tr>
</tbody>
</table>

Table 4.32: This is the result of the Spearman test for OPEX in each group of difference in condition.

### 4.5.2 Trends in aggregate data

In this section follows six scatter plots that show the trends of some key metrics during the time period HYDRA has been used. These results should not be interpreted as consequences of results from HYDRA, but can be used as support to tell if the information gained from the model is used in the correct way. Every plot has a line that represents the least squares estimate of the linear trend in the data.

#### 4.5.2.1 Unplanned spill and lost margin

In Figure 4.16a is a scatter plot showing the trend of the spill for the period 2006 to 2015. There seems like the unplanned spill have increased the last ten years. However, the last data point is twice as large as the others and thus could be seen as an outlier. This point influences the slope of the trend line to be larger than it really is.
Figure 4.16b shows the trend in unplanned lost margin for the period 2006 to 2015. Since the lost margin is calculated from the unplanned spill the two plots are highly correlated, but due to the changes in the price of electricity they look different. Hence it is possible to see the same trend for lost margin as for the unplanned spill, however the slope is not as steep.

Figure 4.16: Scatter plots of spill and lost margin against time.

4.5.2.2 Availability, interruptions and mean condition of units

The availability of the hydropower plants is one of the most important metrics. In Figure 4.17a is the trend of the mean availability of all units for the period 2005 until 2015. The availability is stable around 95%.

In Figure 4.17b the trend of total number of hours with interruptions due to production-related incidents is presented. The linear estimate shows a downward trend, with less number of hours with interruptions. This can also be seen in the total number of failures in Figure 4.17c.

Figure 4.17d shows how the mean condition of the units has changed during the period 2004 to 2015. Here it is possible to see a positive trend where the mean condition have improved during the time period.
(a) Availability without planned interruptions against time in years of the period 2005 to 2015.

(b) Unplanned lost margin against time in years of the period 2005 to 2015.

(c) Mean number of failures of units against time in years for the period 2004 to 2015.

(d) Mean condition of units against time in years for the period 2004 to 2015.

Figure 4.17: Plots of availability, interruptions, failures and mean condition against time.
Chapter 5

Conclusions and recommendations

5.1 Conclusions

The analysis of HYDRA was divided into five parts:

- How well can Vattenfall Hydro assess the condition of the power plants?
- Has Vattenfall Hydro chosen the right set of adverse events for the model?
- Do the probabilities in the risk model reflect reality in a correct manner?
- Does the model successfully predict the risks in the power plants?
- What have the consequences and trends been during the period?

The results show that HYDRA as a model is not performing well. It is hard to say if this is due to that the the model is incorrect or that the input to the model is not assessed in a correct manner.

Analysis of the input, the assessment of the condition of the plants, show that measures which are expected to correlate with the condition does not. However, it is not possible to say that Vattenfall are bad in assessing the status based on this. These power plant are reliable systems, which mean that there are rarely any larger incidents. Hence the condition of one unit can be considered as bad but it can still perform better then a unit that is considered as being in good condition.

The fact that there are very few incidents of larger scale also affect the analysis of determining if the adverse events used in HYDRA to happen. What
can be concluded is that incidents related to important parts do happen, but it is quantitatively impossible to tell if these incidents are due to hazards or just interruptions.

When there are so few events it is also hard to make a proper analysis of the probabilities which are used in HYDRA. As the results show the probabilities in HYDRA are in a wider interval than the calculated probabilities. However, the fact that all calculated probabilities are somewhat the same, makes one suspect that the probabilities in HYDRA are inaccurate.

The fact that it was hard finding insurance accidents that could be discovered in HYDRA may say that HYDRA is doing its job. However, looking into individual events, the facts show that the opposite is possible. It should be taken into consideration that HYDRA is not built to tell the probability of a hazard. HYDRA was built to support decisions about renewals and when these should be planned, and to make it simpler to compare the risks of each unit.

Actions taken in the power plants have some effects, where among other things, the number of failures for units and the number of hours with interruptions have decreased. There is also a small but positive correlation between the change in condition of units between two years and the OPEX. This implies that OPEX can be used to improve the condition of the power plants.

5.2 Recommendations

The recommendations to Vattenfall Hydro based on the results of this thesis, are:

- If HYDRA is used forward, it is important to determine a process to report accidents that are related to HYDRA.
- If a new model that can be evaluated is introduced, consider what data could be of interest for the evaluation and start collecting it from the beginning of usage.
- To make it easier to do analyses it is important to keep interesting data in one location where everyone knows where it is.
- Important to have a mutual view of what a hazard is.
Bibliography


