Data-Driven Methods for Reliability Evaluation of Power Cables in Smart Distribution Grids

Hassan Mashad Nemati

Supervisors:
Sławomir Nowaczyk, Anita Sant’Anna and Antanas Verikas
Abstract

This research aims to develop data-driven methods that automatically exploit historical data in smart distribution grids for reliability evaluation, i.e., analyzing frequency of failures, and modeling components’ lifetime. The results enable power distribution companies to change from reactive maintenance to predictive maintenance by deriving benefits from historical data. In particular, the data is exploited for two purposes: (a) failure pattern discovery, and (b) reliability evaluation of power cables.

To analyze failure characteristics it is important to discover which failures share common features, e.g., if there are any types of failures that happen mostly in certain parts of the grid or at certain times. This analysis provides information about correlation between different features and identifying the most vulnerable components. In this case, we applied statistical analysis and association rules to discover failure patterns. Furthermore, we propose an easy-to-understand visualization of the correlations between different factors representing failures by using an approximated Bayesian network. We show that the Bayesian Network constructed based on the interesting rules of two items is a good approximation of the real dataset.

The main focus of reliability evaluation is on failure rate estimation and reliability ranking. In case of power cables, the limited amount of recorded events makes it difficult to perform failure rate modeling, i.e., estimating the function that describes changes in the rate of failure depending on age. Therefore, we propose a method for interpreting the results of goodness-of-fit measures with confidence intervals, estimated using synthetic data.

To perform reliability ranking of power cables, in addition to the age of cables, we consider other factors. Then, we use the Cox proportional hazard model (PHM) to assess the impact of the factors and calculate the failure rate of each individual cable. In reliability evaluation, it is important to consider the fact that power cables are repairable components. We show that the conclusions about different factors in PHM and cables ranking will be misleading if one considers the cables as non-repairable components.

The developed methods of (a) are applied on data from Halmstad Energi öch Miljö (HEM Nät), Öresundskraft, Göteborg Energy, and Växjö Energy, four different distribution system operators in Sweden. The developed methods of (b) are applied on data from HEM Nät.
I would like to express my sincere gratitude to my supervisors, Associate Prof. Sławomir Nowaczyk and Dr. Anita Sant’Anna. Your knowledge and insightful comments have been invaluable. I also would like to say thank to my advisors, at Halmstad Energi och Miljö (HEM Nät), Alexander Örning and Peter Addicksson for their support and inspiring cooperation.

Many thanks to Prof. Antanas Verikas, for providing a pleasant working environment in the Department of Intelligent Systems and Digital Design. Many thanks to Associate Prof. Stefan Byttner for his great support and provision of helpful suggestions particularly in the PhD education. I also would like to thank the support committee members, Prof. Ulf Johansson and Associate Prof. Robert Bass for providing valuable feedback on my research.

I would like to give special thanks to my friends and colleagues. Thanks for helping me find ways to solve problems, and for the many helpful discussions. Thanks to Pablo del Moral for reviewing the thesis and his comments during the writing process.

Finally, I would like to say thanks to my family. Thanks for their love, guidance and support through my life. Special thanks to my brother Hossein for his guidance during my education and research.

*This project is funded by Knowledge Foundation and HEM Nät.*
Contents

1 Introduction .............................................. 1
   1.1 Challenges ........................................... 3
   1.2 Research Questions .................................. 3
   1.3 Research Contributions .............................. 4
   1.4 List of Appended Publications ...................... 5

2 Background ............................................. 7
   2.1 Probabilistic Reliability Indices .................... 7
      2.1.1 Data Collection ................................. 8
      2.1.2 System Characteristics ......................... 9
      2.1.3 Statistical Method ............................. 10
   2.2 Interruption indices ................................. 12

3 Methodology and Results ............................... 15
   3.1 Fault analysis ....................................... 15
      3.1.1 Method .......................................... 16
      3.1.2 Results ......................................... 17
      3.1.3 Future work ..................................... 18
   3.2 Reliability analysis ................................. 22
      3.2.1 Method .......................................... 22
      3.2.2 Results ......................................... 23
      3.2.3 Future work ..................................... 26
   3.3 Evaluation and planning ............................. 26
      3.3.1 Method .......................................... 26
      3.3.2 Results ......................................... 26
      3.3.3 Future work ..................................... 26

4 Summary of papers ....................................... 29
   4.1 Paper B - Bayesian Network representation of meaningful patterns in
electricity distribution grids .......................... 29
4.3  Paper C - Reliability Evaluation of Power Cables Considering Restoration Characteristics ......................................... 30

5  Conclusion and Perspectives ........................................ 33
  5.1  Conclusion ......................................................... 33
  5.2  Future work ........................................................ 34

References ............................................................... 35

A  Paper A ................................................................. 41
B  Paper B ................................................................. 51
C  Paper C ................................................................. 59
List of Figures

1.1 Input data and outcomes of the thesis. ................................. 2

2.1 Bathtub curve of typical failure rate for components. ................. 8
2.2 Important aspects in reliability evaluation of power components. ... 8
2.3 Schematic plot for events and censoring in feeder lines. .............. 9

3.1 The procedure of the proposed data-driven method for reliability eval-
uation. ............................................................................. 15
3.2 The Bayesian Network representing association rule (A ⇒ B) .... .... 17
3.3 Approximated Bayesian Network representation using different thresh-
old values. These figures show how the complexity of the network
varies depending on the thresholds for confidence and support. ......... 19
3.4 From the historical failures in the network information system dataset,
859 pairwise association rules are generated corresponding to 58 nods.
This figure shows the relation between the complexity of the approx-
imated Bayesian Network and the thresholds for confidence and support.
The darker the area, the higher the thresholds and the smaller
number of connections. ....................................................... 20
3.5 Bayesian Network representation constructed based on the event his-
tory of the power quality measurements system database. ............. 21
3.6 Empirical failure rate per kilometer as function of age, for high volt-
age PILC cables. ............................................................ 24
3.7 Six different failure rate models, fitted to the empirical data after re-
moving the incorrect failures. ............................................ 24
3.8 Tradeoff between the number of customers and the average annual
consumption for the first 20% cables in the case scenario 3 ranking list. 27

5.1 Frame work of the planned activities. .................................... 34
List of Tables

3.1 Goodness-of-fit measurement by using MSE test ............... 25
3.2 Part of the ranking results for HV cables ................ 25
3.3 Three Faulty HV feeder lines and their corresponding rank ...... 27
Chapter 1
Introduction

The reliability of electric power grids (generation, transmission, and distribution) is critically important for both utilities (providers/companies) and customers. Industries, infrastructure, and citizens rely on electric power, and power outages can have disastrous effects. Furthermore, many governing bodies are continuously increasing requirements put on distribution companies concerning the acceptable number and duration of power outages.

Electric power distribution grids, which are the final stages in the delivery of electric power to the end users, are addressed as the most vulnerable sector in power grids [1, 2, 3, 4]. Aging infrastructure, poor design, high exposure to environmental conditions, and irregular electricity usage are some of the common factors causing failures in distribution grids.

Smart distribution grids (SDGs) are designed to be the next generation of power distribution grids. The term “smart” in SDG implies that the grid has the capability to perform self-monitoring and auto-balancing, detect overloads, re-route power, and prevent outages with minimal human intervention [5, 6, 7]. Furthermore, Advanced Metering Infrastructure (AMI), distributed throughout the grid, allows for continuous collection of data. Improving the reliability of electric power delivered to the end users is one of the important objectives of employing smart grid technologies.

Although SDGs provide reliability improvement mechanisms, they require significant additional operational automation to achieve their full promise [8]. For example, SDGs can automatically detect and resolve outages caused by opening and closing interruption devices. However, several outages caused by a failed component need manual work to be precisely localized and fixed [9]. This manual work is usually very costly and time consuming.

To mitigate the number and severity of power outages caused by faulty components, there is a need to switch from reactive maintenance (repair after failures occur) to predictive maintenance (repair before failures occur). The goal in predictive maintenance is to estimate the condition of a system or a component and perform maintenance accordingly. Applying predictive maintenance can prevent catastrophic equipment failures [10].
CHAPTER 1. INTRODUCTION

Power cables (underground cables and overhead lines) are responsible for carrying electrical current over short and long distances. They are one of the fundamental elements in power grids. These cables are heavily affected by ionization, as well as thermal and mechanical stresses [11]. Cable failures usually create long outages. In case of underground cables, both pinpointing and repairing faults are very costly and time consuming due to the difficulty in accessing them.

In SDGs, the majority of reliability improvement techniques are devoted to deviation detection based on real-time streaming data. However, large amounts of historical data related to measurement readings, previous faults, repairs, and manufacturer information are also available but rarely used for predictive maintenance. Mining and analyzing these historical data enables us to design reliability evaluation methods that can estimate the life time of power cables. Then, power distribution companies can directly target the most vulnerable cables for inspection and preventive repair actions.

Clearly, it is not cost efficient to suggest all power cables should be replaced every few years, e.g., every 10 years. For the same reason, it is also impractical to put monitoring devices on each cable. Instead, we want to exploit the “cheap”, accessible, and available historical data for evaluating the reliability of power cables.

This research aims to develop data-driven methods that automatically exploit available historical data in SDGs for reliability evaluation of power cables. The main purpose of this project is to enable power distribution companies to change from a reactive (corrective maintenance) mode to predictive mode by deriving benefit from available data. In particular, we want to use historical data for a) failure pattern discovery, and b) reliability evaluation of power cables (see Figure 1.1). We would like to be able to predict which power cables need maintenance and proactively mitigate events that cause outages.

![Figure 1.1: Input data and outcomes of the thesis.](image-url)
1.1 Challenges

Real-world data - Many practical challenges arise when working with real-world data. For example, how the recorded data represent the “actual” failures of a system or component in the grid. Which systems, components, or factors are causing a specific failure, or increasing the probability of an event. How different features describing the failures are connected and how we can represent their connections.

Quantity and quality of data - In SDGs, large amount of historical data are recorded in different ways, but not all of them can be used for reliability analysis of power cables. Among these, we can refer to customers’ electricity consumption data. Consumption data are mainly used for billing purposes and their relation with cable’s reliability is very weak. On the other hand, the history of previous faults is highly relevant, but for power cables, the amount of historical failure information is usually small. This lack of data makes it difficult to perform reliability evaluation of power cables with reasonable confidence.

Impact of multiple factors on reliability evaluation - Our evidence shows that targeting only the oldest cables for replacement is not an optimal strategy for reliability improvement. In fact, cable age is only one factor among many that impact failure rates. Other factors include previous maintenance, geographical position, length, etc. Some of these factors can be captured directly from datasets, although the corresponding information may be imprecise. In this case, the major challenge is related to identifying which factors should be included and estimating how much is the impact of each factor on power cables’ reliability.

Reliability modeling: repairable vs non-repairable - In case of failure in power cables, usually the faulty point will be replaced by a new segment and the rest of the cable remains intact. Many previous works have been ignoring this repairability characteristic to simplify their model or to deal with limited amount of failure data, despite the fact that their results of failure analysis might be misleading. In reliability evaluation of power cables, it is important to design a robust method that considers the restoration characteristic and limitations associated with repairable systems.

1.2 Research Questions

Based on the aim of the project and previously highlighted challenges, the following research questions are considered in this thesis.

1. How to discover and represent failure pattern from historical failure data of a SDG?

2. How to perform reliability evaluation with confidence bounds while dealing with limited historical failure data?

3. How to design a robust method for reliability evaluation considering the impact of the repairability characteristic of power cables and selection of different factors?
CHAPTER 1. INTRODUCTION

1.3 Research Contributions

The main contributions of this thesis are summarized as follows:

1. Proposing an easy-to-understand visualization of the correlations between different features representing failures by using an approximated Bayesian network (Paper B). Patterns are considered as the correlation between failure and other features such as season, weekday, time, and outage duration. The outcomes of failure pattern discovery can be used for identifying the most vulnerable components (e.g. underground cables) or factors that are important for further analysis. The cause of the identified correlations, the level of the impact of correlated features on failure rate, and ways to limit different failures are some of the topics that can be investigated in the future. The generality of the method is evaluated by applying the method on historical data from other distribution companies (see Chapter 3.1.2).

2. Proposing a methodology for power cables lifetime modeling with confidence intervals to deal with limited failure data (Paper A). We investigated five different models estimating the probability of failures for in-service underground cables. In many practical cases, the amount of data available is very limited, and it is difficult to know how much confidence one should have in the goodness-of-fit results. Therefore, we focused on a methodology for evaluating how well different models fit the historical data and represent the results by confidence bounds. In this analysis only the age of the cables is considered (not the impact of additional factors).

3. Demonstrating the importance of considering the reparability characteristic of power cables on reliability estimation (Paper C). In particular we compared three case scenarios depending on how to consider power cables and their failures: as nonrepairable components, as repairable but decommissioned after the last failure, and as repairable components which survive until censoring time. For power cables, the first and second scenarios are incorrect but often used, and we showed that conclusions about reliability analysis will be misleading if they are used.

4. Developing a method for ranking repairable power cables based on the impact of different factors (Paper C). We used the Cox proportional hazard model (PHM) to assess the impact of different factors and calculate the failure rate of each individual cable. Then we ranked cables based on their failure rate. The method is applied considering the restoration characteristic of power cables. Currently we are doing on-site testing on the highly ranked cables to validate the outcomes of the method. Additional factors, by consulting with experts, are needed to be included.
1.4 List of Appended Publications

The appended publications in this thesis are listed in the following:


In general, reliability is defined as the probability of a system or component performing adequately for the period of time under the intended conditions [12]. The reliability of power grids is considered traditionally regarding two aspects: adequacy and security [13, 14].

Adequacy is defined as the ability of the power grid to supply power electricity and energy requirements for the customers. Security is defined as the ability of the power grid to withstand specific sudden disturbances such as an unanticipated failure in the grid. When a power system in the grid satisfies the security criterion, it is said to be, e.g. “N-1 secure”, meaning that it could lose any one of its N components and continue operating. Similarly, if a power system is “N-k secure” it means that no consumer would loose electricity even if k components were suddenly disconnected [15, 16].

Traditional approaches for reliability evaluation (N-k criteria) are deterministic. They do not take into account uncertainties. Within the deterministic framework, the outage event with the worst consequence sets the system’s reliability and events with lower consequence but higher likelihood may be missed [17]. On the other hand, probabilistic indices directly reflect various parameters that can have impact on power systems’ reliability. They are generally more flexible than deterministic indices but also more complex [16]. In this thesis the probabilistic indices are considered.

2.1 Probabilistic Reliability Indices

The probabilistic indices for reliability evaluation of power systems are usually expressed by terms such as lifetime, failure probability, cumulative failure distribution, and failure rate. Failure rate (FR) is the frequency of failure within a time interval [18, 19]. Failure rate function, \( \lambda(t) \), is a function that describes changes in the rate of failure depending on time. Figure 2.1 shows a commonly used model that represents the failure rate function known as the bathtub curve [20]. The model begins with a high FR (infant mortality), followed by fairly constant FR (useful life). Finally, the FR increases again as the component reaches the end of its life (wear-out).
CHAPTER 2. BACKGROUND

Figure 2.1: Bathtub curve of typical failure rate for components.

In order to perform reliability evaluation, some important aspects should be considered including how to obtain required data, which statistical analysis should be chosen, and what are the constraints and limitations of reliability evaluation due to the specific characteristic of power components (see Figure 2.2).

Figure 2.2: Important aspects in reliability evaluation of power components.

2.1.1 Data Collection

One way to collect the required data for reliability analysis of power systems is to perform on-site testing [21, 22]. In this approach, the actual condition of a system can be measured while the system is in operation, based on some suitable and measurable indication of the system or the component deterioration.
2.1. PROBABILISTIC RELIABILITY INDICES

For power systems, on-site testing may damage the insulation of the testing component [23] and therefore laboratory tests are commonly used instead of on-site testing. In laboratory tests [24, 25, 26], a new component first undergoes accelerated aging processes to simulate the condition of aged ones. Then the component deterioration indicators are measured. Hence, instead of measuring the actual status of the in-service component, the condition of the component will be estimated based on off-site (laboratory) tests.

In case of power cables, both on-site testing and laboratory testing methods are costly and complex processes. Moreover, the condition of power cables can also be estimated in a cheaper way by analyzing historical data. In general, utility companies keep records of historical data such as previous events and inventory data (manufacturer information) which can be used for reliability evaluation of power cables [27, 28, 29, 30, 31].

One of the problems with analyzing historical information is the limited amount of failure data, which makes it difficult to estimate the reliability of power cables with reasonable confidence. To mitigate this problem, some previous works [18, 19] have used expert knowledge to manually refine the estimated lifetime of power cables. These approaches, however, are error prone. In addition, such approaches cannot be easily generalized from one grid to another.

If every feeder line is under observation until a failure, the reliability measures may be estimated simply by computing the fraction of lines surviving at each age. However, some feeder lines have not failed by the end of the experiment (data collection). In this case, the data is called “right censored” [32]. These censored events affect the reliability analysis and should be considered.

![Figure 2.3: Schematic plot for events and censoring in feeder lines.](image)

2.1.2 System Characteristics

In order to perform reliability analysis, one must consider the nature of power systems and the limitations of the reliability evaluation methods. In reliability evaluation, there is a crucial difference between the statistical treatment of repairable systems and non-repairable systems. A repairable system or component can be restored to
CHAPTER 2. BACKGROUND

satisfactory operation after a failure by repair actions. A non-repairable system or component is removed permanently (replaced with a new system or component) after a failure. Ascher and Feingold in [33] and Zapata et.al in [34] discussed some common misconceptions about modeling of repairable systems that may lead to wrong conclusions.

One of the important misconceptions mentioned in [34] is “the use of life model for a repairable component”. According to this paper, the life model refers to the occurrence of one and only one failure that “kills” a non-repairable component. Therefore, it is incorrect to apply life model to a repairable component as it can withstand several failures. Another concept discussed in this paper is the misleading idea that a stationary random process which has constant statistical properties (expected value and variance) over time can represent a non-stationary random process. Thus, “constant failure rate” has been applied as a rule of thumb for any type of components, forgetting the fact that this can be applied only for non-repairable components.

The replacement of an underground power cable is very costly and it is not economically efficient to change the entire cable after a failure. Therefore, in case of failure, only the faulty point will be replaced by a new segment and the rest of the cable stays untouched. The repairability characteristic of power cables allows to keep them in service even after a failure. In fact, as long as the frequency of failures in a specific cable is not high (tradeoff between the cost of multiple repairs and replacing the entire cable), power companies tend to keep the cable in service. Furthermore, power cables after a failure and repair are usually as-bad-as-old, i.e., the repair after each failure does not materially change the condition of the entire cable. In this case, even though we may repair a cable, the age of the cable does not change to zero but stays the same as before.

2.1.3 Statistical Method

In a very broad way, failure data can be evaluated statistically using either parametric methods, or nonparametric methods [20]. Parametric methods make assumptions about the underlying population from which the data are obtained. On the other hand, the non-parametric methods, which are sometimes called “distribution-free” methods, do not assume any particular family for the distribution of the data [35].

In both parametric and nonparametric methods, the failure processes are described as random events. These events are then considered as random variables that can have a continuous or discrete characteristic [36].

If components are considered as non-repairable, the FR or hazard rate (HR) function are usually used to estimate the remaining useful life of the components [20, 37, 38]. The failure time is a random variable T described by a single time to failure. The order of failure times does not matter, i.e. the random variables T are not chronologically ordered. In this case, FR is the relative rate of failure of components surviving until time T (conditional).

If components are considered as repairable, stochastic point process (SPP), renewal process (RP) model, or reliability growth analysis are usually used to estimate
the expected number of events over time [39, 40, 41, 42, 43, 44, 45, 46, 47]. If the failure time $T$ represents the time between successive failures, it is called inter-arrival time. Here it is assumed that the repair action materially changes the condition of the component (the condition of the component after repair is “as-good-as-new”). But, if the repair action does not materially change the condition of the component (the condition of the component after repair is “as-bad-as-old”), the time of failure compared to time 0 represents the failure time $T$. For repairable systems, the rate of occurrence of failure (ROCOF) and mean time between failure (MTBF) are usually used to represent the expected number of cumulative failures at each time stamp. ROCOF is the absolute rate at which system failures occur (unconditional).

The following measures can be used for reliability evaluation [48] of non-repairable components.

**Probability function:** the probability that any randomly chosen component fails during time $t$ to $t + \Delta t$

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t}$$ (2.1)

**Cumulative distribution function:** the probability that any randomly chosen component fails within the interval $(0, t]$

$$F(t) = P(T \leq t) = \int_0^t f(u)\,du$$ for $t > 0$ (2.2)

**Reliability function:** also known as survival function, the probability that any randomly chosen component does not fail within the interval $[0, t]$

$$R(t) = P(T > t) = 1 - F(t) = 1 - \int_0^t f(u)\,du$$ for $t > 0$ (2.3)

**Failure rate function:** the probability that an observed component (the component has not failed yet) fails during time $t$ to $t + \Delta t$

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{R(t) - R(t + \Delta t)}{\Delta t \cdot R(t)} = \frac{f(t)}{R(t)}$$ (2.4)

**Parametric methods**
In parametric methods, depending on the type of the random variable (continuous or discrete), a distribution model is fitted to the data. With different goodness-of-fit tests the model parameters can be estimated. The amount of available data influences the confidence bounds of the performed analysis.

For power system components the Weibull model is commonly used to fit the failure rate data points [49]. The failure rate function of the Weibull model with shape parameter $\beta > 0$ and scale parameter $\eta > 0$ is:
\[ \lambda(t) = \frac{\beta}{\eta} \left( \frac{t}{\eta} \right)^{\beta-1} \]  
\hspace{1cm} (2.5)

Other reliability measures such as probability density function \( f(t) \), and cumulative distribution function \( F(t) \) can be calculated using the following formulas.

\[ f(t) = \frac{\beta}{\eta} \left( \frac{t}{\eta} \right)^{\beta-1} \cdot \exp \left[ - \left( \frac{t}{\eta} \right)^{\beta} \right] \]  
\hspace{1cm} (2.6)

\[ F(t) = 1 - \exp \left[ - \left( \frac{t}{\eta} \right)^{\beta} \right] \]  
\hspace{1cm} (2.7)

Nonparametric methods

Nonparametric models make no assumption about the underlying distribution model, meaning that the distribution of a components’ life time is unknown. Kaplan-Meier and Nelson-Aalen estimators are two examples of nonparametric methods. These estimators are used for calculating survival function \( \hat{S}(t) \) and cumulative FR function \( \hat{\Lambda}(t) \) by the following equations [50]. Here the censoring in data is also considered.

\[ \hat{S}(t_i) = \prod_{j=1}^{i} \frac{n_j - d_j}{n_j} \]  
\hspace{1cm} (2.8)

where \( d_j \) is the number of events at time \( t_j \) and \( n_j \) is the number of subjects “at risk”. Note that the estimator \( \hat{S}(t_i) \) drops only at times when a failure has been observed, not at times when censoring occurs.

\[ \hat{\Lambda}(t_i) = \sum_{j=1}^{i} \frac{d_j}{n_j} \]  
\hspace{1cm} (2.9)

Intuitively, this expression is estimating the failure at each distinct time of event \( t_j \) as the ratio of the number of failures \( d_j \) to the number of components “at risk” \( n_j \). Therefore \( \hat{\Lambda}(t_i) \) estimator is an increasing right-continuous step function with increments \( \frac{d_j}{n_j} \) at the observed event time. The components that are censored are not counted as “at risk”.

2.2 Interruption indices

Interruption indices can give additional information compared to probabilistic and traditional deterministic indices. Interruption indices are developed to represent the “customer-oriented”, “load-oriented”, and “energy-oriented” characteristics of failures [51] such as the number of disconnected customers due to a power outage or the duration of power outage for specific customers. Some of the widely used indices are
listed in the following.

**SAIFI:** the system average interruption frequency index is a measure of the average number of interruptions in a year

\[
\text{SAIFI} = \frac{\text{total number of customer interruptions}}{\text{total number of customers}} \tag{2.10}
\]

**SAIDI:** the system average interruption duration index is the average outage time in a year for each customer in the network

\[
\text{SAIDI} = \frac{\text{total number of customer hours of interruptions}}{\text{total number of customers}} \tag{2.11}
\]

**CAIDI:** the customer average interruption duration index is the average duration of an interruption experienced by the customers interrupted

\[
\text{CAIDI} = \frac{\text{total number of customer hours of interruptions}}{\text{total number of customer interruptions}} \tag{2.12}
\]

Each country has defined specific rules for electricity distribution operators to improve the reliability of their SDG by limiting the interruption indices. Some of the Swedish regulations for acceptable limit of outages are described in the following [52]. More information about Swedish law can be found in [53].

From 2006, all the distribution system operators in Sweden have to report risk and vulnerability analysis regarding to the reliability of their electricity grid. The risk analysis has to contain action plans of how the reliability shall be improved. The customers should be compensated with 12.5% of the annual tariff after 12 hours of outage (minimum amount of 100 €). This compensation increases if the outage time is more than 12 hours. From 2008, information about extensive outages has to be reported to Energy Markets Inspectorate (EI), an independent authority responsible for regulating tariff levels, within 14 days. The following are considered as extensive outages [52]:

- Outage longer than 24 hours and involves more than 1,000 customers or 25% of the customers.
- Outage longer than 12 hours and involves more than 10,000 customers or 50% of the customers.
- Outage longer than 2 hours and involves more than 100,000 customers.

From 2011 interruptions above 24 hours are not tolerated.
Chapter 3
Methodology and Results

This chapter presents our proposed data-driven method for reliability evaluation of
distribution power grids and in particular power cables, by using historical data. This
methodology contains three steps: fault analysis, reliability analysis, and evaluation
of the results. These steps are illustrated in Figure 3.1 and explained in the following.

![Figure 3.1: The procedure of the proposed data-driven method for reliability evaluation.](image)

3.1 Fault analysis

The available historical data sources for this project are network information system
dpPower, supervisory control and data acquisition (SCADA) system, power quality
measurements system, and smart meters (SM) from a distribution system operator
(DSO) in the south of Sweden (HEM Nät). Among these, we mainly utilize data from
the network information system. This data source contains information about grid
architecture and topology (detailed digital maps), systems and components inventory, maintenance history, electricity customers, etc. Moreover, some information about events, cessation of a system or components’ ability to perform its required function, is stored in this database.

3.1.1 Method

In the fault analysis step, a preliminary investigation of event history data from the network information system database is used to identify the “high-priority” components. Components that fail more frequently, cause long outages, are more costly to repair, or affect large number of customers can be considered as high-priority components. Furthermore, potential factors that have impact on failures and the correlations of these factors with each other are specified in this step.

In general, utility companies keep records of previous faults which contain features describing the event. To analyze failures, it is important to discover which failures have common features, e.g., if there are any types of failures that happen mostly in certain parts of the grid or at a certain time. One approach to discover these correlations is employing association rules [54, 55, 56, 57]. Association rules are based on the frequency of the co-occurrence of features and conditional dependency between them.

The objective of mining association rules is to find the most frequently occurring combinations of features. Let \( I = \{I_1, I_2, \ldots, I_m\} \) be a set of features (items). An association rule is an implication of the form \( A \Rightarrow B \), where \( A \subset I \), \( B \subset I \), and \( A, B \) are disjoint itemsets, i.e. \( A \cap B = \emptyset \). In this case the itemset \( A = \{a_1, a_2, \ldots\} \) is the prior and the itemset \( B = \{b_1, b_2, \ldots\} \) is the posterior of the rule. Now assume that \( X = \{x_1, x_2, \ldots, x_n\} \) is a set of random variables representing the list of observations (failures) in a dataset. Each observation \( x_i \) in the dataset \( X \) may or may not contain a specific item, e.g., \( x_1 = \{I_1, I_2, I_5\} \) only contains items \( I_1, I_2, I_5 \).

The interestingness of an association rule \( A \Rightarrow B \) is often expressed in terms of support, confidence, and lift.

- The support of a rule is the percentage of observations in the dataset that contain both \( A \) and \( B \).
- The confidence of a rule is the percentage of examples containing \( A \) that also contain \( B \). In other words, a fraction that shows how frequently \( B \) occurs among all the observations containing \( A \).
- The lift of a rule is a ratio of the confidence of the rule to the frequency of observations containing \( B \). It is a value between 0 and infinity that measures the deviation of a rule from statistical independence.

The confidence value of each association rule corresponds to the strength of the conditional dependence between features. Therefore, these confidences can be used for automatically building a Bayesian Network.
Bayesian Networks [58, 59, 60] are graphical representation of probabilistic relationships over a set of variables, constructed using probability distribution over a set of variables in a dataset. If we consider features of failure events as probabilistic variables, a Bayesian Network captures the conditional relations between those features over a set of events.

An association rule \( A \Rightarrow B \) can be seen as a connection from one itemset to another. If \( I = \{I_1, I_2, \ldots, I_t, \ldots, I_m\} \) is a set of features such that \( A = \{I_1, I_2, \ldots, I_t\} \) and \( B = \{I_{t+1}, \ldots, I_m\} \), the Bayesian Network representation for all the connections between feature set \( A \) and \( B \) is shown in Figure 3.2, where items in set \( A \) are parent nodes and items in set \( B \) are child nodes.

\[
P(I_1, I_2, \ldots, I_t, \ldots, I_m) = \prod_{i=1}^{t} P(I_i) \prod_{j=t+1}^{m} P(I_j|I_1, I_2, \ldots, I_t) \quad (3.1)
\]

Each of the terms \( P(I_j|I_1, I_2, \ldots, I_t) \) corresponds to the confidence of the rule \( ((I_1, I_2, \ldots, I_t) \Rightarrow (I_j)) \).

### 3.1.2 Results

According to historical data at HEM Nät, the most common failure is caused by “Fabrication fault” with a frequency of 34.59%. According to the experts at the company, any fault that is related to aging of components is recorded under label “Fabrication fault”. The “affected component” statistics show that underground feeder cables are one of the most common faulty components in the grid, with a frequency of 26.94%. Among these feeder cables, breakdown in PILC cables and joints (because of aging) is the most common cause of failure.

The association rules with “high” support and confidence, which also have lift greater than 1, are considered as interesting rules. One example of the identified “interesting rules” is Fusebreak \( \Rightarrow \) UngPillar. According to its confidence, we can interpret this rule as: the probability that an underground cable pillar is the affected component knowing that the cause of failure is fuse break is 79.061%.
Finally, we use association rules with two items to construct Bayesian Network (approximated Bayesian Network). For this purpose the lists of priors and posteriors of each rule correspond to the network nodes, and the confidence of the rule (conditional dependency) corresponds to the connections between nodes. Figure 3.3 illustrates three networks constructed from the association rules but with different thresholds for confidence and support. In Paper B we show that the Bayesian Network constructed based on the interesting rules of two items is a good approximation of the real dataset and it can be used for calculating conditional probabilities of association rules for more than two items.

The methodology for discovering failure patterns and failure statistics, presented in Paper B, was also applied on historical data from Öresundskraft, Göteborg Energy, and Växjö Energy. These companies are DSOs which are located in Sweden. The results show that the most vulnerable component in these distribution grids is either underground cable or overhead line. Furthermore, correlated features with failures are identified. Interesting failure patterns are discovered using association rules and represented by Bayesian Networks.

3.1.3 Future work

The features that are highly correlated with failures of underground cables can be used as the indication of important factors. Thus, these important factors can be used for further analysis to estimate how the variation of different factors can change the cables’ failure rate.

One of the constrains in automatically constructing approximated Bayesian Network from association rules is the manual threshold setting for confidence and support. For HEM Nät data, the fully connected failure network contains 859 connections between all the nodes from different categories. However, we assume that some of the items are independent or the dependencies can be neglected, since they are very weak (confidence and support smaller than a certain threshold). In fact, the selected thresholds for confidence and support specify whether to consider a rule as an “interesting rule” or not. This manual setting can be tuned by considering the complexity and accuracy of the network. Figure 3.4 shows the relation between the complexity of the Bayesian Network (as the number of connections between nodes) and the thresholds for confidence and support. The darker the area, the higher the thresholds and the smaller number of connections. If the thresholds are too small, the number of connection in the network is high and consequently interpreting the result will be difficult. In the other case if the thresholds are too high, the number of connections is too small to capture the “interesting” correlations.

The procedure of constructing Bayesian Networks can also be used for other type of databases. Figure 3.5 shows the result of Bayesian Network, constructed based on one of the datasets in the power quality measurements system database. This dataset contains information about recorded events occurred in one main station with two transformers. In this figure, the thicker connections represent high confidence (above 80%). Several observation can be made from this network. For example, most of the
3.1. FAULT ANALYSIS

Figure 3.3: Approximated Bayesian Network representation using different threshold values. These figures show how the complexity of the network varies depending on the thresholds for confidence and support.
events of type “sag” happened during Saturday, Sunday, Hour night (between 23:00 and 07:00), or Summer. The transformer T31 has more recorded events during Autumn and the transformer T32, during Summer. This representation of the events provides an easy-to-understand visualization of the most relevant failure patterns.

Furthermore, we have reconfigured a number (more than 1000) of SM in the grid to collect more data, e.g., power quality, frequency, and total harmonics. One important type of data that we are collecting from SMs is *alarm data*. These alarms are indicators of disturbances such as sag, swell, which occur to the customers. Applying the fault analysis on this data, and more specifically, finding the correlation between the alarms and cable failures is one of the planned activities for continuing this research.
Figure 3.5: Bayesian Network representation constructed based on the event history of the power quality measurements system database.
The results enable distribution companies to discover failure patterns and accordingly mitigate the conditions that increase the probability of failures.

3.2 Reliability analysis

The main focus of this step is on failure rate modeling of underground cables (the identified high-priority component from the previous step) based on historical data.

3.2.1 Method

Our methodology for reliability analysis contains four major steps: pre-processing, modeling baseline failure rate function, estimating the impact of relevant factors on failure rate function, and ranking components based on their failure rate.

To perform reliability analysis we need to identify previous events in underground cables. In our case, the historical events database could not be directly linked to the feeder line information, because the two use different asset identifiers. Therefore, we use the assumption that short cable sections (length smaller than 20 meters) in any given line are artifacts of previous repairs. The failure is assumed to have taken place in the year of the installation of the short cable section, and to take place in the oldest cable within this line. Therefore the year difference between the installation of the oldest cable and the short section is considered as the age of the cable when the failure is occurred. Those assumptions are not fully accurate, but we have confirmed through discussions with domain experts that they are realistic. Based on these assumptions, some feeder lines are incorrectly linked to a few extra failure events. For example, several short cable sections may be installed because of upgrading a sub station, and not because of a failure. In order to eliminate these extra failures we exploit the sub station maintenance history, which contains information about the previous maintenance carried out on the sub stations and the connected feeder lines. We determined that between January 1929 and December 2015, the number of events that occurred on high voltage (HV) and low voltage (LV) lines are 213 and 331, respectively.

Baseline failure rate is a function that estimates the variation of failure rate over age ($\lambda_0(t)$). In estimating the baseline failure rate, only the age of cables is considered as a factor that impacts failure rate. In Paper A, a methodology for estimating the baseline failure rate while dealing with limited amount of failure data is explained. It is important to note that, while goodness-of-fit (GOF) results can be compared directly, it is often difficult to properly interpret the results, especially when the data are of limited quantity. Therefore we propose a method for interpreting the results of GOF measures with confidence intervals, estimated using synthetic data. Five different models are used to fit with the empirical data. For each model, a number (e.g. 100) of synthetic data sets are generated by drawing random points from a normal distribution with mean equal to the failure function at each age and variance computed from the empirical failure rate data points. Then the GOF of the synthetic data are computed in compare with all other models to determine how well a data generated
from one model can be fitted by another model. These comparisons help us to draw conclusions about how well each model fits the empirical data points.

One way to estimate the impact of additional factors on failure rate function is to use a Proportional Hazard Model (PHM). PHM is a statistical regression model which was first introduced by Cox in 1972 [61]. PHM is based on the assumption that the failure rate of a system or component consists of two multiplicative coefficients: the baseline failure rate \( \lambda_0 \), and an exponential function, including the effect of factors \( \exp (\beta \cdot X) \).

\[
\lambda(t, X_t) = \lambda_0(t) \cdot \exp (\beta \cdot X_t)
\]  

where \( \lambda_0(t) \) is the baseline failure rate that is dependent on time \( t \), \( X_t \) is a row vector representing the factors at time \( t \), and \( \beta \) is a column vector representing the regression parameters. The vector of factors \( X_t \) can be time-dependent or time-independent [62]. The coefficient vector \( \beta = [\beta_1, \beta_2, \ldots, \beta_n] \) is a vector of regression parameters. \( \beta_i \) could be positive (positive correlation with the failure), or negative (negative correlation with the failure). If \( \beta_i = 0 \) then the expression \( \exp (\beta_i \cdot X) \) for the \( i \)th factor is equal to 1, meaning that the factor does not affect the failure rate. An estimate of the \( \beta_i \) values, without making any assumption about the baseline failure rate, can be calculated by a maximum likelihood function [63].

After modeling the baseline failure rate and measuring the influence of different factors, we calculate the failure rate for each individual feeder line. The feeder lines are then ranked from the highest failure rate to lowest. The feeder lines with higher rank indicate greater vulnerability and the need for remedial actions.

### 3.2.2 Results

In Paper A we investigate five different models (linear, piecewise linear, exponential, constant, and piecewise constant) to model baseline failure rate and evaluate how well, each model fits empirical failure rate data points. We interpret the results by comparing the obtained GOF measures with expected GOF and confidence intervals, estimated using synthetic data. Observe that, we do not specifically consider the “infant mortality” period in this analysis.

The result of calculating empirical failure rates at each age, for high voltage PILC cables, is shown in Figure 3.6. In this figure the failure rate at ages between 32 and 35 have very high values (unexpected values compare to other data points). Our investigations show that some feeder lines are incorrectly linked to a few extra failure events because of upgrading a sub-station, and not because of a failure. Therefore, we eliminate these extra failures for further analysis.

After removing the “incorrect” failures, we fit six different models with the empirical failure data points. The result of fitting these models are shown in Figure 3.7 and the GOF measures presented in Table 3.1. According to the table, the constant and piecewise constant models (models A and B) are statistically different from the
rest of the models. Furthermore, there is no statistically significant difference between GOF of the data points, neither the empirical nor synthetic, between linear, piecewise linear, exponential, and Weibull models. This indicates that these four models are virtually identical. Therefore, for our failure rate data points there is no difference between choosing either of these four models.

As mentioned in the Chapter 2, it is important to determine if a component is repairable or non-repairable. To show how this would impact the results of cable ranking and conclusions about different factors, we consider three case scenarios. These scenarios are defined based on considering power cables and their failures: as non-repairable components, as repairable but decommissioned after the last failure, and as repairable components which survive until censoring time. In Paper C we show that, when analyzing the long time history of failures for power cables, the first
and second scenarios are incorrect, and conclusions about different factors in PHM and cables ranking will be misleading if they are used.

Table 3.2 shows the ranking results for the 10 highest ranked feeder lines based on case 1. Then the corresponding ranking values and ranking positions for these lines according to case 2 and 3 are added for comparison. Accordingly, for some feeder cables such as H200 and H306 the ranking position is the same in all three cases; some other lines such as H105 have slightly different position; and finally, there are some feeder lines such as H996 which are ranked completely different in positions 10, 17, and 99 for cases 1, 2, and 3 respectively. It is not very important if the line is ranked in position e.g. 4, 5, and 7 in the three lists, but the very high difference between ranks e.g. 10 and 99 (changing from a very high rank to a very low rank) is significant and cannot be neglected.

Table 3.2: Part of the ranking results for HV cables

<table>
<thead>
<tr>
<th>ID</th>
<th>Case 1 Rank</th>
<th>Case 1 FR</th>
<th>Case 2 Rank</th>
<th>Case 2 FR</th>
<th>Case 3 Rank</th>
<th>Case 3 FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>H200</td>
<td>1</td>
<td>0.626</td>
<td>1</td>
<td>0.121</td>
<td>1</td>
<td>0.105</td>
</tr>
<tr>
<td>H306</td>
<td>2</td>
<td>0.035</td>
<td>2</td>
<td>0.051</td>
<td>2</td>
<td>0.037</td>
</tr>
<tr>
<td>H784</td>
<td>3</td>
<td>0.03</td>
<td>3</td>
<td>0.044</td>
<td>5</td>
<td>0.03</td>
</tr>
<tr>
<td>H105</td>
<td>4</td>
<td>0.025</td>
<td>5</td>
<td>0.035</td>
<td>7</td>
<td>0.025</td>
</tr>
<tr>
<td>H843</td>
<td>5</td>
<td>0.024</td>
<td>7</td>
<td>0.031</td>
<td>6</td>
<td>0.029</td>
</tr>
<tr>
<td>H205</td>
<td>6</td>
<td>0.021</td>
<td>6</td>
<td>0.033</td>
<td>8</td>
<td>0.024</td>
</tr>
<tr>
<td>H971</td>
<td>7</td>
<td>0.019</td>
<td>4</td>
<td>0.035</td>
<td>3</td>
<td>0.034</td>
</tr>
<tr>
<td>H206</td>
<td>8</td>
<td>0.018</td>
<td>8</td>
<td>0.029</td>
<td>12</td>
<td>0.022</td>
</tr>
<tr>
<td>H996</td>
<td>9</td>
<td>0.017</td>
<td>13</td>
<td>0.024</td>
<td>49</td>
<td>0.016</td>
</tr>
<tr>
<td>H914</td>
<td>10</td>
<td>0.01</td>
<td>17</td>
<td>0.021</td>
<td>99</td>
<td>0.014</td>
</tr>
</tbody>
</table>
3.2.3 Future work

In addition to the considered factors, there are other factors that may impact the power cables failure rate such as the type of the soil where the cables are buried; or the number/type/energy consumption of customers who are connected to each feeder line. Combining the information acquired from this step and SMs alarm data for improving the reliability ranking is planned for the future.

3.3 Evaluation and planning

The purpose of this step is to evaluate the results of reliability analysis and cables ranking computed based on the three case scenarios.

3.3.1 Method

To evaluate the results of the ranking lists, created based on the three case scenarios, we consider two options. The first option is to search for recent failures in the grid and look into the position of the faulty feeder lines in each list. The other option is to perform some tests such as Time Domain Reflectometry (TDR) on the high-ranked lines.

TDR is a method to localize the faulty part of a feeder line by sending a low-energy signal through the line. The printout of TDR, also known as “trace”, is a graphical representation of the return signal which gives an approximate location of impedance variations. Moreover, for in-service healthy feeder lines it is possible to perform TDR. In this case, the trace localizes the weak points in the line such as weak joints and location of partial discharge. Therefore by performing the TDR method on the high-ranked cables, it is possible to evaluate the performance of the ranking approach.

3.3.2 Results

During the four month period after creating the list, three failures (caused by aging of the cables) have occurred in HV feeder lines. The ranks of these three lines based on the case scenarios are shown in Table 3.3. Accordingly, among 458 HV feeder lines these failures have happened on the lines which are ranked in the first 15% of the case scenario 3 rank list. These three examples are not enough to evaluate the performance of the ranking approach but it gives us some real observation of faulty lines and their position in the ranking lists based on the three cases.

3.3.3 Future work

The TDR measurements are being planned with the company for the autumn of 2017. From the ranking list of case scenario 3, the top 20% of the high ranked feeder lines are selected. In addition, to specify which line should be selected for measurement
Table 3.3: Three Faulty HV feeder lines and their corresponding rank

<table>
<thead>
<tr>
<th></th>
<th>Rank - case 1</th>
<th>Rank - case 2</th>
<th>Rank - case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cable 1</td>
<td>80 (% 17.5)</td>
<td>14 (% 3.1)</td>
<td>44 (% 9.6)</td>
</tr>
<tr>
<td>Cable 2</td>
<td>102 (% 22.3)</td>
<td>50 (% 10.9)</td>
<td>63 (% 13.8)</td>
</tr>
<tr>
<td>Cable 3</td>
<td>76 (% 16.6)</td>
<td>209 (% 45.6)</td>
<td>37 (% 8.1)</td>
</tr>
</tbody>
</table>

and test we consider two other factors: number of customers, and average annual consumption of customers who are connected to each feeder line. Then we used Pareto frontier which is defined as the graphical representation of the tradeoffs within a set of parameters. The result is shown in Figure 3.8. In this figure, the data point correspond to the feeder lines. The bigger the data point, the higher the failure rate. Based on the Pareto frontier (tradeoff between the number of customers and the average annual consumption) and discussion with experts from the company, five feeder lines are selected for testing.

Figure 3.8: Tradeoff between the number of customers and the average annual consumption for the first 20% cables in the case scenario 3 ranking list.
Chapter 4
Summary of papers

4.1 Paper B - Bayesian Network representation of meaningful patterns in electricity distribution grids

In Paper B, three different ways for detecting fault patterns in an electricity distribution grid using history of event data are presented: failure statistics, association rules, and Bayesian networks. We proposed a simplified representation of the association rules by using Bayesian Networks. We show that a small subset of the most interesting rules is enough to obtain a good and sufficiently accurate approximation of the original dataset.

In general, utility companies keep records for previous faults that contain features describing the event such as time, date, cause, faulty component, etc. To analyze failures characteristic it is important to discover which failures have common features, e.g., if there are any types of failures that happen mostly in certain parts of the grid or at certain times. Primary evaluation of the historical failure is used for analyzing the frequency of occurrence for each failure in an electricity distribution grid. Statistical analysis and association rules are applied to discover correlation between the features. Association rules are based on the frequency of the co-occurrence of features and conditional dependency between them. Their interestingness is often expressed in terms of probability. If we consider features in recorded events as probabilistic variables, a Bayesian Network captures the conditional relations between those features over a set of events.

The results provide a clear and practical representation of features associated with events that can be used by maintenance staff at electricity distribution companies. The outcomes of the proposed method for discovering failure pattern facilitate the choice of considering the most vulnerable components, e.g., underground cables or important factors for further analysis.

In this paper, we focus on the methodology for evaluating how well different models fit empirical failure rate while dealing with limited amount of failure data. We analyze five different models to estimate the relationship between the age and failure rate in underground high voltage cables. As is common in this domain, the amount of failure data is limited, and it is difficult to know how much confidence should one have in the GOF results. The proposed methodology is based on interpreting the results by comparing the obtained GOF measures with expected GOF and confidence intervals, estimated using synthetic data.

In addition to commonly used models, we also consider constant and piecewise constant models. For each model, a number of synthetic data sets are generated by drawing random points from a normal distributions with mean equal to the failure function at each age and variance computed from the empirical data points.

According to the result of GOF tests, the linear, piecewise linear, and exponential models do not show significant difference. This indicates that those three models are virtually identical when they are used to fit empirical failure rate data. On the other hand, the piecewise constant model fits the failure rates better, in a statistically significant way, than other models.

As it is described in Chapter 3.2.2, some feeder lines are incorrectly linked to a few extra failure events because of upgrading a sub-station. These “incorrect” failures affect the results of GOF and evaluating the models. After identifying these failures, removing them, and fitting different models with the empirical failure data points, we observe that the constant and piecewise constant models are statistically different from the rest of the models. Furthermore, for our data points there is no statistically significant difference between choosing either of linear, piecewise linear, exponential, or Weibull models.

4.3 Paper C - Reliability Evaluation of Power Cables Considering Restoration Characteristics

In this paper we show that it is important to consider the repairability characteristics of power cables and choose the reliability analysis which is designed for repairable systems. We demonstrate that the methods which estimate the time-to-the-first failure (for non-repairable components) may lead to incorrect conclusions about reliability of power cables.

We use Cox proportional hazard model (PHM) to assess the impact of different factors and calculate the failure rate of each individual cable. After modeling the PHM baseline and the influence of different factors, we calculate failure rate for each individual feeder line, and rank them from the highest failure rate to the lowest. In particular we compare three case scenarios depending on how to consider power cables and their failures: as non-repairable components, as repairable but decommissioned
after the last failure, and as repairable components which survive until censoring
time. In principle, for power cables, the first and second scenarios are incorrect, and
we show that conclusions about different factors in PHM and cables ranking will be
misleading if they are used.

The results show that the significance level of the factors in PHM is different
considering each case scenarios. Furthermore, the variation between the ranking lists
shows that the case scenarios produce different outcomes for reliability ranking. This
variations between the lists are not negligible.

By ranking the components importance to system reliability, the awareness of the
grid status can be improved and actions can be taken to reduce the risk.
Chapter 5
Conclusion and Perspectives

5.1 Conclusion

In smart grids, the majority of reliability improvement techniques are devoted to deviation detection based on real-time streaming data. However, large amounts of data related to measurement readings, previous faults, repairs, and manufacturer information are also recorded but rarely used for predictive maintenance. Mining and analyzing these historical data enables us to evaluate reliability and estimate power components life time. Then, power companies can directly target the most vulnerable components for inspection and preventive repair actions.

In this thesis and appended papers we present data-driven methods that automatically exploit available historical data in SDGs for reliability evaluation. In particular, the historical data is used for two purposes: (a) failure pattern discovery, and (b) reliability evaluation of power cables. Power cables are one of the fundamental elements in power grids and cable failures usually create long outages. In case of underground cables, both pinpointing and repairing faults are very costly and time consuming.

The developed methods of (a) are applied on data from Halmstad Energi och Miljö (HEM Nät), Öresundskraft, Göteborg Energy, and Växjö Energy, four different distribution system operators in Sweden. The developed methods of (b) are applied on data from HEM Nät.

The main contributions of this thesis are summarized in the following.

- Proposing an easy-to-understand visualization of the correlations between different features representing failures by using a Bayesian network.
- Proposing a methodology for power cables lifetime modeling with confidence intervals to deal with limited failure data.
- Demonstrating the importance of considering the reparability characteristic of power cables on reliability estimation.
- Developing a method for ranking repairable power cables based on the impact of different factors.
5.2 Future work

The future work is grouped into three categories (see Figure 5.1). The first is related to testing cables, evaluating the reliability ranking method, and modifying the reliability analysis. Additional factors such as type of the soil where the cables are buried and the number/type/energy consumption of customers who are connected to each feeder line will be considered.

The second category is related to exploiting the data collected from SMs and power quality measurements system databases for deviation detection of power components. In this case, we will investigate on failure pattern discovery and deviation detection methods to find correlations between the alarms and failures in the grid.

The third category is maintenance planning. In SDGs, planning for maintenance is an essential part in achieving “effective” maintenance and consequently improving reliability. Based on the reliability analysis, condition of in-service power cables are estimated so that maintenance can be conducted on cables with higher failure rank and higher importance, e.g., number or type of customers. We will investigate methods and techniques for maintenance planning based on the relationship between failure rank, importance, and cost.

Figure 5.1: Frame work of the planned activities.
References


REFERENCES


REFERENCES


[47] E. L. Mariut and E. Helerea. Enhancing reliability for medium voltage underground power lines. 7th International Symposium on Advanced Topics in Electrical Engineering (ATEE), pages 1–6, 2011. (Cited on page 11.)


Appendix A

Paper A
Reliability Evaluation of Underground Power Cables with Probabilistic Models

Nemati, Hassan M., Anita Sant’Anna, and Sławomir Nowaczyk

Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) 2015
Reliability Evaluation of Underground Power Cables with Probabilistic Models

Hassan M. Nemati, Anita Sant’Anna, Sławomir Nowaczyk

Abstract—Underground power cables are one of the fundamental elements in power grids, but also one of the more difficult ones to monitor. Those cables are heavily affected by ionization, as well as thermal and mechanical stresses. At the same time, both pinpointing and repairing faults is very costly and time consuming. This has caused many power distribution companies to search for ways of predicting cable failures based on available historical data.

In this paper, we investigate five different models estimating the probability of failures for in-service underground cables. In particular, we focus on a methodology for evaluating how well different models fit the historical data. In many practical cases, the amount of data available is very limited, and it is difficult to know how much confidence should one have in the goodness-of-fit results.

We use two goodness-of-fit measures, a commonly used one based on mean square error and a new one based on calculating the probability of generating the data from a given model. The corresponding results for a real data set can then be interpreted by comparing against confidence intervals obtained from synthetic data generated according to different models.

Our results show that the goodness-of-fit of several commonly used failure rate models, such as linear, piecewise linear and exponential, are virtually identical. In addition, they do not explain the data as well as a new model we introduce: piecewise constant.

I. INTRODUCTION

Electric power transmission and distribution networks consist of different types of cables, some of which have been installed more than 50 years ago, and some are newly added to the network. The major problem with these power cables is the lack of efficient condition monitoring methods [14].

Power outages, i.e., the unavailability of electricity supply due to faults, have many undesirable effects and are a high cost to the society as a whole. Loss of production, cost of repair, and customers’ dissatisfaction are some of the important factors to be considered when analyzing the impact of outages. For institutions like hospitals, airports, and train stations, power outages can be disastrous.

There are many different reasons for power outages. According to a study by the Edison Electric Institute [10], 70 percent of power outages in the USA are weather related phenomena such as lightning, rain, snow, ice, etc. Another 11 percent of outages are caused by animals, such as birds, coming into contact with power lines. To reduce the impact of such incidents, many power electric companies are moving towards underground transmission and distribution lines. However, underground cables may also cause outages, most commonly due to insulation degradation and ruptures in conductors.

One drawback of underground cables is that the procedure for finding the exact place of failure is harder, since no visual inspection can be performed. In addition, even when a fault is localized, the process of digging the ground to reach the cable, and also repairing the cable, is more difficult and requires more skill than for aerial cables.

Many governing bodies are continuously increasing requirements put on distribution companies concerning the acceptable number and duration of power outages. In addition, in many areas of life, society is more and more relying on electrical power. Consequently, there is a great need for better methods to determine the condition of the in-service underground cables and their remaining useful life. In particular, it is important that those methods are cost effective.

In this paper, we analyze five different models to estimate the relationship between the age and failure rate in underground high voltage cables. In addition to commonly used models (linear, piecewise linear, and exponential), we also consider constant and piecewise constant models. In particular, we focus on the methodology for evaluating how well different models fit the data. As is common in this domain, the amount of data we have available is very limited, and it is difficult to know how much confidence should one have in the goodness-of-fit results.

We calculate the empirical failure rates based on real data of over fifty years of historical faults from a small European city. The data comes from historical databases at Halmstad Energi och Miljö (HEM Näät), one of the Swedish electricity distribution companies.

The remaining of this paper is structured as follows. Background and related work is presented in section 2. In section 3 we explain the proposed model evaluation methodology, and we describe our experiments and results in section 4. We summarize our contribution and discuss future work in section 5.

II. BACKGROUND AND RELATED WORKS

A mathematical model that represents the current condition of a cable is known as the state of the cable [13]. The state represents the condition of the cable at a given point in time. Owing to the fact that the cables are laid under the ground, their current state is not directly observable. Depending on the amount of available information, one can estimate the state in different ways, using different models. Clearly, if the information about the cables increases, the
representing model becomes more precise. However, there is a tradeoff between the cost of collecting additional data and the benefits such data would provide.

There are mainly two methods for condition assessment of underground cables. The first is measuring the cables’ condition by using different types of diagnostic and stress test analysis such as partial discharge (PD) and dielectric losses. The second is mining historical information such as age of the cables, and previous failures.

The condition of power cables can be measured in two ways: using on-site testing [6], [7], [9] or laboratory testing [16]. On-site testing is performed directly on the in-service cables. In the laboratory testing, first, a new cable undergoes accelerated aging processes to simulate the condition of aged cables, which are then analyzed. In both of these methods the amount of PD, oil analysis, and bulk properties of insulation, e.g. tanδ measurement, are used to determine the cables condition. The tanδ measurement is a diagnostic test conducted on cables’ insulation to measure their deterioration. In fact, the tanδ measurement is used as the loss factor of the insulation material which will increase during the aging process. The assessment of the in-service cables should be performed every 3-5 years and the results classify the investigated cables into different categories based on which future maintenance can be performed. Both of these measurements are very costly and complex processes.

The historical data analysis is usually performed in one of the two ways. The first is based on Crow-AMSAA and reliability growth model [1], [2], [8], [14]. Based on the time duration between each recorded failure in the system, historical failures are modeled using a Weibull distribution. This Weibull model is then used to estimate the time to the next failure, usually in the whole system, i.e., for all underground cables, without any distinction between aged and new cables. In other words, all the cables are considered to be in the same condition, regardless of their age, type, and other factors.

In the second historical data analysis method, in addition to the previous failures, other information such as age, and insulation condition are used to model failure rate [11], [12], [18]. Bloom et al. [3], [4] used historical data for age and number of previous failures as “observable condition”; and experts’ judgment for insulation degradation condition, environmental stressor, and effect of the previous failures as “unobservable conditions”. By using the historical data and the experts’ knowledge they modeled the changes in cables’ condition probabilistically, i.e., given the current state of a cable, what is the probability of different cable states in the future. Of all the factors used in their work, only age and historical failure rate are extracted from actual data, and all the rest of the information is based on the experts’ judgment.

The failure rate model is usually used for estimating the expected number of future failures. One important aspect is that future failures are influenced by the replacement strategy employed, which is one of the possible solutions for electric power companies to reduce the number of outages. Replacement actions, also known as rejuvenation, is the procedure of replacing the old and faulty parts with new cables. There has been some research analyzing how the replacement of old cables reduces the number of expected failures and improves reliability, for example [11] and [12], however, the majority of work in the field does not take rejuvenation into account.

In general, there are three types of underground cables widely used in distribution power grids [5]:

- Oil-Filled cable
- Paper Insulated Lead Cover cable (PILC)
- Cross-linked Polyethylene cable (XLPE)

Before development of XLPE cables in 1993, PILC cables were the most common installed underground power cables [15]. Their estimated expected lifetime is declared to be around 40 years [17], but they have been used for more than that in many transmission and distribution grids. In these grids, the problem of degradation of underground cables due to aging is becoming more and more severe.

The old Paper Insulated Lead Cover (PILC) cables, which are of main concern in this study, are heavily affected by a number of factors such as ionization, thermal breakdown as well as electrical and mechanical stresses [5]. Since the paper insulation is made of cellulose, the quality of the insulation degrades over time and causes more frequent breakdowns. One way to decrease the corrosion speed and cable fragility is to fill the paper insulation with oil.

There are several important factors that accelerate the aging process in PILC cables. The ones most commonly mentioned in the literature are cyclic overloading, thermal breakdown, PD, irregular load pattern, direct or indirect spiking, inadequate depth in the ground, and very low temperature.

Cable joints, which are part of the underground cables, can also cause outages in the network. The jointing is the act of reconstructing two cables to become one. It is used when a longer cable is needed or when a part of an old cable is replaced with a new cable. A joint is usually the weakest part of an underground cable and it is affected by three types of stressors: thermal, electrical, and mechanical stress. Mechanical stress and water ingress are the main causes of failures in cable joints [5]. The fault in the joints might affect the conductor, insulation, or sheath. The sheath of the joints get corroded due to overloading and the chemicals present in the soil over a period of time. This increases the chance of moisture seepage into the joint, which subsequently causes failure.

In this work, we only use available historical data to compute failure rate. This approach is not as accurate as performing direct measurements on individual cables, but is often preferred in practice since mining the available data to find a model is significantly cheaper than performing laboratory or field tests.

III. METHODOLOGY

It is well known that by analyzing historical information of cables inventory, it is possible to predict the future failures in
cables with some degree of accuracy. One common example is modeling the failure rate for a particular type of cables. We use historical data from a small European city to estimate the parameters of the model. This model can then be used to predict future faults for different cables.

In particular, in this paper we focus on the failure rate for PILC underground cables at a certain age. Note that there are several other factors affecting failure rate variation in cables, such as number of joints, history of previous failures, environmental stressors, usage patterns, manufacturer and cable type, etc. Here, however, we only consider the age and the number of historical faults to estimate failure rate.

To estimate failure rate we need to have access to historical databases containing information such as installation year, date of previous failures, and the age of the cable at the time of failure. Furthermore, to calculate the proportion of faulty cables over total cables, we need to know the total length of all the in-service cables during each year.

The process of calculating the failure rate, estimating model parameters, and finally, evaluation of the results is described below, as shown in Figure 1.

A. Pre-processing

Due to the requirements explained above and the available databases, we have selected cable inventory data set. This data set contains historical information about both the in-service and destroyed cables that have been installed since 1908 in Halmstad power distribution grid. Each cable is described with a unique ID and the transmission line to which it belongs, as well as additional information such as insulation type, conductor size, installation year, length, etc. In this work, we only analyze in-service high-voltage PILC cables.

A transmission line between two cable boxes consists of a number of cables. According to the data set, the total number of high-voltage transmission lines containing PILC cable is about 500.

The cables in a line may have different installation years. We assume that the initial installation year of the line is the earliest installation year among all cables in the group.

In addition to length of in-service cables, we require information about past failures. In our case the historical failure database could not be directly linked to the cable information, since the two use different asset identifiers. Therefore, to identify past failures, we use the assumption that short cables in any given line are artifacts of previous repairs. Therefore, we consider each cable of length smaller than 20 meters to correspond to a failure in the line. The failure is assumed to have taken place in the year of the installation of the short cable, and to take place in the oldest cable within this line. Those assumptions are not fully accurate, but we have confirmed, through discussions with domain experts, that they are realistic.

B. Failure rate estimation

Failure rate is the frequency with which a system or component fails within a given unit of time. This definition can be naturally extended to a population of systems, for example a network of cables. In this work we consider the number of failures per year per kilometer. The general equation for the empirical failure rate is:

\[ FR = \frac{N}{L}, \]

where \( N \) is the number of failures in a year and \( L \) is the total length of in-service cables.

There are many factors that influence the failure rate, however, in this work we only focus on cable age (understood as the number of years between installation of the cable and the time of the failure). It is a well-known fact that the likelihood of failure changes with age. Therefore, we express the empirical failure rate for underground cables at age \( \alpha \) as the total number of failure that happened to cables at age \( \alpha \), denoted \( N(\alpha) \), divided by the total length of cables that were in-service at age \( \alpha \), denoted \( L(\alpha) \):

\[ FR(\alpha) = \frac{N(\alpha)}{L(\alpha)}. \]

Among several factors affecting failure rate, we only consider the factors that can be estimated from the historical databases we have access to: installation year of each cable (age), length, voltage class (high voltage or low voltage), and failure history: number of failures, and age at time of failure.

C. Modeling and parameter estimation

A failure function, \( \lambda(\alpha) \), is a function that describes changes in failure rate depending on age. Figure 2 shows a commonly used model that represents the failure function known as the bathtub curve [11]. The model begins with a high failure rate (infant mortality), followed by fairly constant failure rate (useful life). Finally, the failure rate increases again as the component reaches the end of its life (wear-out).

![Fig. 1. Overview of the model creation and evaluation process.](image)
When discussing power cables, we are particularly interested in modeling the “wear-out” time and the effects of aging process on failure rate. It is commonly believed that the failure rate increases as cables get older.

Our goal is to find an appropriate failure function $\lambda(\alpha)$. To this end, we investigate five different models and evaluate how well, the empirical failure rates fit each model. Observe that we do not specifically consider the “infant mortality” period in this analysis.

We have decided to perform experiments using five different failure functions. The three commonly used models in statistical analysis are linear, piecewise linear, and exponential. In addition to these, we have also investigated constant and piecewise constant models.

**Constant:** This model is described by a constant line with the failure rate equal to $\lambda(\alpha) = \mu$, where $\mu$ is the mean failure rate value of all the empirical data points $FR(\alpha)$.

**Piecewise constant:** This model is constructed by two constant lines at different values, $\mu_1$ and $\mu_2$, where $\mu_1$ is the mean failure rate before $T_{pwc}$ and $\mu_2$ is the mean failure rate after $T_{pwc}$.

$$\lambda(\alpha) = \begin{cases} 
\mu_1 & \text{if } T_{pwc} \leq \alpha \\
\mu_2 & \text{if } T_{pwc} > \alpha 
\end{cases}$$

**Linear:** The linear model is specified by a linear function with two parameters: slope $m_l$ and intercept $b_l$. In this model, the increment of failure rate between two consecutive time points is constant.

$$\lambda(\alpha) = m_l(\alpha) + b_l$$

**Piecewise linear:** This model represent the failure rate to be constant at the beginning up to age $T_{pwl}$, and then failure rate grows linearly with slope $m_{pwl}$. Therefore, the function is specified by three parameters, the constant failure rate $b_{pwl}$, the time which failure rate starts to increase linearly $T_{pwl}$, and the slope $m_{pwl}$ of the line.

$$\lambda(\alpha) = \begin{cases} 
b_{pwl} & \text{if } T_{pwl} \leq \alpha \\
m_{pwl} \cdot (\alpha - T_{pwl}) + b_{pwl} & \text{if } T_{pwl} > \alpha 
\end{cases}$$

**Exponential:** this distribution is described by the function:

$$\lambda(\alpha) = \beta \cdot e^{\beta \cdot \alpha}$$

For each model, the corresponding parameters are calculated by Levenberg-Marquardt optimization algorithm implemented in Python `scipy` library, minimizing the mean square error.

After parameter estimation, we need to evaluate how well do the empirical data points fit each model. This can be done by using different goodness-of-fit measures.

### D. GOF evaluation

In this study we employ two goodness-of-fit measures; the first is based on calculating the probability of generating the data from a given model (PGD); the second is based on mean square error between the data and the model (MSE).

In the PGD measure, for each age, the value of the failure function $\lambda(\alpha)$ at that age is considered to be the mean value of a normal distribution. The variance of this normal distribution is computed from the empirical data points. At each age, the cumulative probability function is used to calculate the probability that a given data point belongs to the normal distribution centered around the failure function. Finally, the calculated probabilities for each age are multiplied together to give the value of GOF for that model. The higher this probability is, the better the data points fit the model.

However, the resulting numbers are very small and difficult to analyze, and thus we use the negative logarithm (base 10) of those values to make them easier to interpret. Therefore, the lower the value of the GOF, the better the empirical failure rates fit the model under consideration.

$$GOF_{PGD} = -\log_{10} \prod_{\alpha} P(x \leq FR_{\alpha}|FR_{\alpha} \in X, \sim (\mu = \lambda(\alpha), \sigma^2))$$

The MSE GOF measure is the sum of squared differences between each data point and the value of the failure function at corresponding age. Also in this case, the lower the GOF value the better the data points fit the model under consideration.

$$GOF_{MSE} = \frac{1}{n} \sum_{\alpha} (FR(\alpha) - \lambda(\alpha))^2$$

where $n$ is the number of data points.

Finally, it is important to note that, while GOF results can be compared directly, it is often difficult to properly interpret the results, especially when the data is of limited quantity (and also quality) and it does not fit any of the models perfectly. Therefore, we propose a way to interpret the results by comparing the obtained GOF measures with expected GOF and confidence intervals, estimated using synthetic data.
Fig. 3. Empirical failure rate per kilometer as function of age, for high voltage PILC cables.

For each model, a number of synthetic data sets are generated by drawing random points from a normal distribution with mean equal to the failure function at each age and variance computed from the empirical data points. The synthetic data sets should have the same number of points as the empirical data points. The PGD and MSE GOF are computed between each synthetic data set and the corresponding model. Confidence intervals are derived based on the variance of the GOF values. The GOF of the synthetic data sets generated by one model are also compared to all other models in order to determine how well a data set generated from model A fits model B. These comparisons will help us draw conclusions about the how well the empirical data points fit each of the proposed models.

IV. RESULTS AND DISCUSSION

The result of calculating empirical failure rates at each age, for high voltage PILC cables, is shown in Figure 3. The horizontal axis represents the cable age at time of failure and the vertical axis represents the failure rate \( \lambda \) (per kilometer).

By comparing this result with Figure 2 it is possible to extract three lifetime phases. The empirical data starts with higher failure rates at ages 1-6, the “infant mortality” period. It then continues with a period of low and fairly constant rates during ages 7-19, the “useful life”. And finally, the higher failure rates start again from age 20, the “wear-out” phase. However, there are also some differences from the bathtub curve, the most clear ones being the peak at ages around 30 years, and the shape of the wear-out phase.

The parameters for constant, piecewise constant, linear, piecewise linear, and exponential models were estimated from the empirical data. Each resulting model is shown in Figure 4. The resulting parameter for the constant model is \( \mu = 0.052 \), and for the piecewise constant model are \( \mu_1 = 0.023 \), \( \mu_2 = 0.082 \), and \( T_{pwc} = 30 \). The parameters for the linear model are \( m_1 = 0.0013 \), and \( b_1 = 0.0128 \). For the piecewise linear \( b_{pwl} = 0.0231 \), \( m_{pwl} = 0.00147 \), and \( T_{pwl} = 0.00695 \). For the exponential model, the parameter \( \beta \) is equal to 0.0254.

To compare the results of GOF between different models, first we generated 100 synthetic data sets based on each model, and then measured the PGD and MSE between each randomly generated data set and all the models. In Figure 5, one randomly generated data set is shown for each model. Then, for each group of 100 generated data sets, we found the mean value of all calculated GOF to all models and the corresponding 95 percent confidence interval.

We performed the PGD and MSE tests for all combination of synthetic data sets and models. In this case, data sets A, B, C, D, and E are the 100 randomly generated data sets from constant, piecewise constant, linear, piecewise linear, and exponential models respectively. The results of GOF tests based on PGD and MSE are presented in Table I and Table II. For example, the result of PGD GOF test of the data generated from constant model (A) with respect to the linear model (C) is \( 43.8960 \pm 0.6270 \).

From the GOF results presented in Table I and Table II, several observations can be made, as follows.

As expected, the best GOF results are obtained when the data set is compared to the model which generated it. For example, Data A fits model A better than any other model. These correspond to the diagonal entries in Table I and Table II.

The results of GOF measurements from fitting each generated synthetic data with the same model (diagonal of the tables) does not show any statistically significant differences. This verifies that performing this type of comparison between synthetic data and models is systematically correct, i.e., the result of comparing model A with synthetic data A is as good as comparing model B with synthetic data B.

Except the constant model (model A) which is statistically very different from the rest of the models, the result of pairwise comparison between a GOF test in synthetic data generated by a model but fitting with another model, and a result of GOF test in the other combination of this two models, is not significantly different. For example, GOF between data B and model D, is not significantly different than the GOF between data D and model B.

There is no statistically significant difference between GOF of the data points, neither the empirical nor synthetic, between linear, piecewise linear, and exponential models. That is, the GOF results are within the respective confidence interval obtained from the synthetic data. This indicates that those three models are virtually identical.

Nonetheless, the real data seems to fit the piecewise constant model better than the other models. This suggests that the failure rate could be modeled by two constant lines; low failure rate up to age 30 and higher failure rate after that. This does not confirm the assumption that the failure rate increases monotonically as a function of age. This observation is quite surprising, and we believe it deserves
further analysis in the future.

This result might be caused by several factors. First, the very high values of failure rate at ages between 32 and 35 years affect other models more than the piecewise constant model. Second, the input data set and the in-use information is not enough to uniquely and with confidence identify the best model. Therefore, other information should also be taken into the account. Third, we did not considered the affect of repair and replacement of cables on failure rate estimation. In fact, the process of rejuvenation of the underground cables prevents the failure rates from becoming too high, especially after experiencing number of failures (in our case after age of 40 years or so).

V. CONCLUSION AND FUTURE WORK

In this paper we have presented some of the characteristics of power grid cables, especially PILC underground cables, which are used in many power transmission and distribution networks. We have also discussed the main challenges regarding fault prediction for these cables.

We have introduced five different probabilistic models for predicting failure rate depending on cable age, and evaluated how well does each of these models fit the real-world, historical fault data. We have employed two different goodness-of-fit measurements, one based on mean square error and one based on probability of generating the data.

In order to compare the GOF measures between various models, a new methodology is presented. The GOF test results are interpreted by generating 100 synthetic data sets for each model, and estimating the corresponding confidence intervals. Then, pairwise comparisons are performed between each model and synthetic data sets.

According to the result of GOF from PGD and MSE tests, the linear, piecewise linear, and exponential models do not show significant difference. On the other hand, the piecewise constant model fits the failure rates better, in a statistically significant way, than other models.

This result was quite surprising, since we expected that the failure rate to be an increasing function of age. This could be explained by the fact that the faulty cable sections are

<table>
<thead>
<tr>
<th>(log10 [10^4])</th>
<th>Data A</th>
<th>Data B</th>
<th>Data C</th>
<th>Data D</th>
<th>Data E</th>
<th>Real Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant (Model A)</td>
<td>38.8769</td>
<td>±0.5529</td>
<td>43.4234</td>
<td>±0.8101</td>
<td>42.0662</td>
<td>±0.8031</td>
</tr>
<tr>
<td>Linear (Model C)</td>
<td>46.9332</td>
<td>±0.9202</td>
<td>38.8123</td>
<td>±0.5135</td>
<td>39.8805</td>
<td>±0.5102</td>
</tr>
<tr>
<td>Exponential (Model E)</td>
<td>43.8565</td>
<td>±0.6775</td>
<td>39.2477</td>
<td>±0.5134</td>
<td>37.0529</td>
<td>±0.8006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(log10 [10^4])</th>
<th>Data A</th>
<th>Data B</th>
<th>Data C</th>
<th>Data D</th>
<th>Data E</th>
<th>Real Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant (Model A)</td>
<td>1.2495</td>
<td>±0.0455</td>
<td>1.6574</td>
<td>±0.0583</td>
<td>1.5392</td>
<td>±0.0633</td>
</tr>
<tr>
<td>Linear (Model C)</td>
<td>1.7007</td>
<td>±0.0779</td>
<td>1.1214</td>
<td>±0.0497</td>
<td>1.1822</td>
<td>±0.0439</td>
</tr>
<tr>
<td>Exponential (Model E)</td>
<td>1.7557</td>
<td>±0.0929</td>
<td>1.2952</td>
<td>±0.0697</td>
<td>1.1794</td>
<td>±0.0609</td>
</tr>
</tbody>
</table>

TABLE II

GOODNESS-OF-FIT MEASUREMENT BY USING MSE TEST

<table>
<thead>
<tr>
<th>(log10 [10^4])</th>
<th>Data A</th>
<th>Data B</th>
<th>Data C</th>
<th>Data D</th>
<th>Data E</th>
<th>Real Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant (Model A)</td>
<td>1.0353</td>
<td>±0.0353</td>
<td>1.1819</td>
<td>±0.0353</td>
<td>1.1278</td>
<td>±0.0353</td>
</tr>
<tr>
<td>Linear (Model C)</td>
<td>1.1278</td>
<td>±0.0353</td>
<td>1.1819</td>
<td>±0.0353</td>
<td>1.1278</td>
<td>±0.0353</td>
</tr>
</tbody>
</table>
continuously replaced by new cables. In fact, the replacement strategy in the underground cables is something we plan to look into in the future in more detail.

In this work we have only considered the failure rate based on the age and the total number of previous failures. However, from the available data set, we can obtain the effects of failure rate based on other factors such as the number of joints, history of previous failures, geographical location, etc. For example, we can cluster cables based on the number of joints per kilometer, and then calculate the failure rate for cables at each cluster. Therefore, by adding more information to the failure rate estimation we can have a better interpretation of the cables failure rate variation over age.

The probabilistic model can also be updated by considering additional information such as load patterns, temperature, and effects of replacement. By exploiting useful information one can determine the condition of in-service equipment, and better plan the scheduling maintenance. Consequently, instead of unplanned outages, power distribution companies can have planned outages, which are shorter and less disruptive.

REFERENCES


Appendix B

Paper B
Bayesian Network representation of meaningful patterns in electricity distribution grids

Nemati, Hassan M., Anita Sant’Anna, and Sławomir Nowaczyk

Energy Conference (ENERGYCON), 2016 IEEE International. IEEE
2016
Bayesian Network Representation of Meaningful Patterns in Electricity Distribution Grids

Hassan M. Nemati, Anita Sant'Anna, Slawomir Nowaczyk
Center for Applied Intelligent Systems Research (CAISR)
Halmstad University
Halmstad, Sweden
hassan.nemati@hh.se, anita.santanna@hh.se, slawomir.nowaczyk@hh.se

Abstract—The diversity of components in electricity distribution grids makes it impossible, or at least very expensive, to deploy monitoring and fault diagnostics to every individual element. Therefore, power distribution companies are looking for cheap and reliable approaches that can help them to estimate the condition of their assets and to predict the when and where the faults may occur.

In this paper we propose a simplified representation of failure patterns within historical faults database, which facilitates visualization of association rules using Bayesian Networks. Our approach is based on exploring the failure history and detecting correlations between different features available in those records. We show that a small subset of the most interesting rules is enough to obtain a good and sufficiently accurate approximation of the original dataset. A Bayesian Network created from those rules can serve as an easy to understand visualization of the most relevant failure patterns. In addition, by varying the threshold values of support and confidence that we consider interesting, we are able to control the tradeoff between accuracy of the model and its complexity in an intuitive way.

Index Terms—Smart Grids, Condition Monitoring, Data Mining, Failure Statistics, Association Rules, Bayesian Networks.

I. INTRODUCTION

Industries, infrastructure, and citizens are relying on power electricity, and consequently electricity outages can have disastrous effects on them. This raises a need for proper strategies for power failure prediction and prevention. One of the common approaches for this purpose is exploiting failures history. In this paper we present three different directions for detecting fault patterns in an electricity distribution grid: Failure Statistics, Association Rules, and Bayesian Networks. The results can be used to design better maintenance strategies to prevent common outages.

Failure statistics is a practical technique for component reliability evaluation which is addressed in many previous works [6]–[8], [10], [11], [13]. In this case, failure probability, failure rate, principal failure causes, events classification, and mean-time-between-failures (MTBF) are the most commonly considered indicators.

In general, the utility companies keep records for previous faults that contain features describing the event. To analyze failures characteristic it is crucial to discover which failures have common features, e.g., if there are any types of failures that happen mostly in certain parts of the network or at certain times. In the literature this analysis is known as discovering patterns in event sequences. One approach is based on finding association rules, and has been addressed by several researchers [1], [3], [5], [14]. Association rules are based on the frequency of the co-occurrence of features and conditional dependency between them. Their interestingness is often expressed in terms of probability.

Bayesian Networks [2], [9], [12] are graphical representations of probabilistic relationships over a set of variables, constructed using probability distribution over a set of variables in a dataset. If we consider features of failure events as probabilistic variables, a Bayesian Network captures the conditional relations between those features over a set of events. In [4], Fauré C. et al. used a five step algorithm to model the frequent association rules using Bayesian Networks. The first step of their algorithm is to create a Bayesian Network based on expert domain knowledge, and then compute association rules for all the combinations of features. The interestingness measure of a rule, derived from the knowledge-driven Bayesian Network, makes it possible to filter out unimportant rules. Tian D. et al. in [12] proposed a Bayesian association rule mining algorithm (BAR) which combines the association rule algorithm with Bayesian Networks. They compute association rules for all the combinations of features and then construct a Bayesian Network. Finally they interpret the result using Bayesian confidence and Bayesian lift using the interestingness of the association rules.

Our approach is based on Tian’s result, however, we do not need to compute the association rules for all combinations of features, which is a very time consuming task. Instead, we construct the Bayesian Network by using only the rules with two features. We show that such a network is still a good approximation of the original dataset, and most of the strong associations can be represented by the joint probability distributions.

In this paper, we explore the real fault history of an electricity distribution company from south of Sweden. Therefore, first we present a simple statistical analysis of historical failures in this grid. Then we present analysis of relations
between features (time, place, the corresponding main-station and sub-station, switchgear, voltage level, cause of the failure, etc.) in the dataset and their co-occurrences. We compute association rules between features, and select the ones with high support and confidence as interesting, since they are describing high statistical correlations. Finally, we propose a Bayesian Network representation of the association rules by using only the rules with two features. We show that most of the strong association rules are conveniently represented by the joint probability distributions of this Bayesian Network, by making some assumptions about the dependency between directed and undirected nodes. This representation provides a simplified visualization of the conditional connections between features and at the same time is much faster to compute than the association rules for all combinations of features.

II. MEANINGFUL PATTERN DISCOVERY

A. Failure Statistics

Primary evaluation of the historical failure is used for analyzing the frequency of occurrence for each failure in an electricity distribution grid. In this case two factors are commonly considered: the probability of occurrence, and the mean-time-between-failures (MTBF). The first factor is calculated as the ratio of the number of each failure over the total number of failures during a specific time interval, and can be used for determining e.g. the most unreliable components. The second factor is the expected number of days between failures of a given type.

B. Mining Association Rules

The objective of mining association rules is to find the most frequently occurring combinations of features. Let \( I = \{I_1, I_2, \ldots, I_m\} \) be a set of features (items). An association rule is an implication of the form \( A \Rightarrow B \), where \( A \subset I \), \( B \subset I \), and \( A, B \) are disjoint itemsets, i.e. \( A \cap B = \emptyset \). In this case the itemset \( A = \{a_1, a_2, \ldots\} \) is the prior and the itemset \( B = \{b_1, b_2, \ldots\} \) is the posterior of the rule. Now assume that \( X = \{x_1, x_2, \ldots, x_n\} \) is a set of random variables representing the list of observations (events) in a dataset. Each observation \( x_i \) in the dataset \( X \) may or may not contain a specific item e.g. \( x_1 = \{I_1, I_2, I_3\} \) only contains items \( I_1, I_2, I_3 \).

We define \( X_A \) as \( x \in X \) that contains items \( A \). In this case, the support of itemset \( A \), represented by \( S(A) \), is the ratio of the cardinality of \( X_A \) over the cardinality of the dataset \( X \)

\[
S(A) = \frac{|X_A|}{|X|} = P(X_A).
\]  

The support of a rule, denoted as \( S(A \Rightarrow B) \), is the percentage of observations in the dataset that contain both \( A \) and \( B \):

\[
S(A \Rightarrow B) = \frac{|X_{A \cup B}|}{|X|} = \frac{|X_A \cap X_B|}{|X|} = P(X_A, X_B). \tag{2}
\]

The confidence of an association rule is the percentage of examples containing \( A \) that also contain \( B \); or, in other words, a fraction that shows how frequently \( B \) occurs among all the observations containing \( A \):

\[
C(A \Rightarrow B) = \frac{S(A \Rightarrow B)}{S(A)} = \frac{P(X_B | X_A)}{P(X_A)}. \tag{3}
\]

The confidence value indicates how reliable the rule is and in this work we use it to measure the interestingness of the pattern.

The lift of an association rule is a ratio of the confidence of the rule to the frequency of observations containing \( B \). It is a value between 0 and infinity that measures the deviation of a rule from statistical independence:

\[
L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{S(B)} = \frac{P(X_A, X_B)}{P(X_A)P(X_B)}. \tag{4}
\]

A lift value smaller than one indicates negative correlation, equal to one indicates no correlation, and greater than one indicate positive correlation between features \( A \) and \( B \) among all the observations.

C. Constructing Bayesian Networks

The confidence value of each association rule corresponds to the strength of the conditional dependence between features, therefore, they can be used for automatically building a Bayesian Network.

An association rule \( A \Rightarrow B \) can be seen as a connection from one itemset to another. If \( I = \{I_1, I_2, \ldots, I_t\} \) is a set of features such that \( A = \{I_1, I_2, \ldots, I_t\} \) and \( B = \{I_{t+1}, \ldots, I_m\} \), the Bayesian Network representation for all the connections between feature set \( A \) and \( B \) is shown in Figure 1, where items in set \( A \) are parent nodes and items in set \( B \) are child nodes.

In general, the joint probability distribution represented by a network can be written as:

\[
P(I_1, I_2, \ldots, I_{t-1}, I_{t+1}, \ldots, I_m) = \prod_{i=1}^{m} P(I_i | parents(I_i))
\]

In our case, where the itemset \( A \) is the parent of itemset \( B \), the joint probability distribution represented by the network can be written as:

\[
P(I_1, I_2, \ldots, I_{t-1}, I_{t+1}, \ldots, I_m) = \prod_{i=1}^{t} P(I_i) \prod_{j=t+1}^{m} P(I_j | I_1, I_2, \ldots, I_t)
\]  

\[ \tag{5} \]

Fig. 1. The Bayesian Network representing association rule \( A \Rightarrow B \)
As shown in equation (2), each of the terms $P(I_1|I_2,I_3,...,I_t)$ corresponds to the confidence of the rule $((I_1, I_2, ..., I_t) \Rightarrow (I_t))$.

### D. Reasoning with Bayesian Networks

We would like to compute the conditional probability of more than two items e.g. $P(X_A|X_B,X_C)$ by using the characteristics of Bayesian Networks and the available connections between set of two items. For this purpose we consider two situations:

- If there is no direct connection between $X_B$ and $X_C$ in the Bayesian Network, we make the simplifying assumption of their independence, which allows us to use formula (6) to approximate $P(X_A|X_B, X_C)$ by the following:

$$P(X_A|X_B, X_C) = P(X_A|X_B)P(X_C) = \frac{P(X_B)P(X_C)}{P(X_B)}$$

This value can be calculated directly from the Bayesian Network, without the need to reference the original data.

- If, on the other hand, there exists at least one direct connection in the Bayesian Network between $X_B$ and $X_C$, the assumption of their independence would lead to too significant errors. In this case we need to consider the two nodes to be dependent. We define the connection with highest confidence as the “primary” connection (without loss of generality we assume that it is from $X_C$ to $X_B$) and approximate $P(X_A|X_B, X_C)$ by the following:

$$P(X_A|X_B, X_C) = P(X_A|X_B)P(X_C) = \frac{P(X_B)P(X_C)}{P(X_B)}$$

In the following section we show that these simplifying assumptions still provide, in practice, good enough approximations of the actual empirical probabilities.

### III. Experimental Results

In this study we consider the Halmstad Energi och Miljö electricity distribution grid (HEM Nät) in the south of Sweden. We use failure history in the grid as our input dataset. This dataset contains information about historical failures for the entire grid during years 2009 until 2015, with a total of 1110 failures. Information such as date, time, the corresponding main- and sub-station, cause of the failure, the faulty component, and outages duration for all the failures in the grid are registered as features.

To discover fault patterns we represent the result using three different methods: failure statistics, association rules, and Bayesian Networks.

#### A. Failure Statistics

The failures in our dataset are grouped into operational, environmental, and unknown failures. The operational failures are caused by a defect in an internal component, the environmental failures are caused by an external factors, and the root cause of failures for unknown faults are unspecified. In Table I the frequency of different types of operational and environmental failures (in percent) and the MTBF from 2009 until the end of 2014 are calculated. According to this table the most common failures during these years are “Fabrication fault” and “Fuse break” with frequency 43.59% and 24.95%, respectively. On the other hand, the most common environmental failure during this period is “Digging” which happened 14.41% of the time. According to the MTBF, the “Fabrication fault” and “Fuse break” occurred in average every 5.7 days and 7.91 days respectively. The failures caused by “Digging” have occurred every 13.69 days.

<table>
<thead>
<tr>
<th>Type of Failure</th>
<th>Cause of Failure</th>
<th>Frequency(%)</th>
<th>MTBF(days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational</td>
<td>Incorrect operation</td>
<td>1.35%</td>
<td>136.88</td>
</tr>
<tr>
<td></td>
<td>Incorrect installation</td>
<td>7.12%</td>
<td>27.72</td>
</tr>
<tr>
<td></td>
<td>Overload</td>
<td>5.59%</td>
<td>35.32</td>
</tr>
<tr>
<td></td>
<td>Lack of maintenance</td>
<td>1.44%</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>4.17%</td>
<td>168.46</td>
</tr>
<tr>
<td>Non-Operational</td>
<td>Digging</td>
<td>14.41%</td>
<td>13.69</td>
</tr>
<tr>
<td></td>
<td>Traffic</td>
<td>3.71%</td>
<td>315.26</td>
</tr>
<tr>
<td></td>
<td>Weather</td>
<td>3.42%</td>
<td>35.63</td>
</tr>
<tr>
<td></td>
<td>Animal</td>
<td>0.72%</td>
<td>273.75</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>0.18%</td>
<td>1095</td>
</tr>
</tbody>
</table>

#### B. Association Rules

Each recorded fault is characterized by different features describing the failure. In order to discover association rules
between these features, we first create a boolean representation of those features. Then association rule mining techniques are used to discover the frequency of co-occurrence of the features and their correlation. In the experiments we have used the following:

- **Season** — Spring (Mar, Apr, May), Summer (Jun, Jul, Aug), Autumn (Sep, Oct, Nov), Winter (Dec, Jan, Feb)
- **Weekday**
- **Hour** — Hour morning (7-12), Hour lunch (12-13), Hour afternoon (13-18), Hour evening (18-22), Hour night (23-7)
- **Main-station** — H2, H3, H4, H7, H8, H10 and Others
- **Outage duration** — less than one hour ($T<1$), between one and two hours ($1<T<2$), between two and three hours ($2<T<3$), between three and four hours ($3<T<4$), greater than 4 hours ($T>4$)
- **Sub-station** — there are 199 sub-stations in the network
- **Cause of failure** — there are 27 types of failures, of which the most important ones are listed in Table I
- **Affected component** — there are 21 types of components, of which the most important ones are listed in Table II

The association rules with high support and confidence, which also have lift greater than 1, are considered as interesting rules. Selected rules are shown in Table III. These rules can be interpreted in the following ways:

**row number 1** — the probability of an outage of duration less than one hour occurring when the affected component is an underground cable pillar is 23.514%. These two are the most common items happening together.

**row number 4** — the probability that an underground cable pillar is the affected component knowing that the cause of failure is fuse break is 79.061%.

**row number 14** — whenever there is a failure in sub-station N78 it is summer, and the probability of seeing this failure again is 0.36%. These two items are highly correlated since the lift is 3.437.

**row number 19** — if there is thunder and it is summer, there is a probability of 0.811% that a failure occurs at main-station H7.

**row number 21** — knowing that there is a digging in areas connected to main station H7 that affects the underground feeder cable, we can expect to have long duration outage (between 2-3 hours) in part of the grid with probability 3.153%.

Some of these rules confirm intuitions our expectations regarding certain types of failures. For example, the rule in row 11 confirms the fact that digging would cause failure in underground feeder cables. Another example is Overload $\Rightarrow$ Winter, with frequency 39 and confidence 62.5%; if we know that the cause of failure is overload, then its a high probability that the season is winter. Similarly, the rule Sunday, Hour morning $\Rightarrow$ 2<T<3 occurs 32 times with confidence 66.672%. It states that if a failure happens on Sunday morning, it takes a long time to be discovered and repaired.

### C. Bayesian Networks

We propose a new approach where association rules with two items are used for constructing a Bayesian Network. For this purpose the lists of priors and posteriors of each rule correspond to the network nodes, and the connections between them correspond to the conditional dependency between random variables. In this section we show that the Bayesian Network constructed based on the interesting rules of two items is a good approximation of the real dataset and it can be used for calculating conditional probabilities of association rules for more than two items. To this end we compute the conditional probabilities of association rules for three features using several examples. Four classes of relations between features and the corresponding connections are shown in Figure 2.

![Fig. 2. Some examples of connected nodes selected from Bayesian Network](image-url)
The fully connected failure network contains $2^{276}$ connections between all the nodes. However, we assume that some of the items are independent or the dependencies can be neglected since they are very weak (confidence smaller than a certain threshold). Considering this assumption the constructed Bayesian Network for our dataset is shown in Figure 3. In this figure, the threshold values are set such that itemsets with frequency higher than or equal to 200 have at least 40% confidence level, itemsets with frequency between 100 to 200 have at least 45% confidence level, itemsets with frequency between 10 to 100 have at least 50% confidence level, itemsets with frequency between 5 to 10 have 80% confidence level, and itemsets with frequency equal to 4 have 100% confidence level. The connections with confidence greater than 80% are shown with bold arrows which corresponds to the availability of strong confidence (more than 80%) between the two nodes.

Based on this network and formula (6) we can calculate the conditional probability of three features when two of them are independent. Examples of such features are (a) and (b) in Figure 2:

$$P(G_{CablePil}|Monday, Hourmorning) = \frac{P(G_{CablePil}|Monday) P(G_{CablePil}|Hourmorning)}{G_{CablePil}} = 50.56\%$$

$$P(G_{CablePil}|Saturday, Hourevening) = \frac{P(G_{CablePil}|Saturday) P(G_{CablePil}|Hourevening)}{P(G_{CablePil})} = 61.404\%$$

Based on this network and formula (7) we can calculate the conditional probability of three features which have direct connections and thus are not independent. Examples of such features are (c) and (d) in Figure 2:

$$P(3<T<4|G_{FeederCa}, Digging) = \frac{P(3<T<4, G_{FeederCa}, Digging)}{P(G_{FeederCa}|Digging) P(Digging)} = 18.429\%$$

$$P(G_{CablePil}|Overload, Winter) = \frac{P(G_{CablePil}, Overload, Winter)}{P(Overload, Winter) P(Overload)} = 51.284\%$$

Although the joint probability value of feature sets with more than three features can be calculated directly from dataset, the selection of interesting itemset and their conditional probability are captured from the important and strong connectivity between nodes in the Bayesian Network (Figure 3).

The comparison between results of computing the above conditional probabilities from the association rules with three features and the Bayesian Network are shown in Table IV. It can be seen that, except for the probability of $P(G_{CablePil}|Saturday, Hourevening)$, the probabilities computed from the Bayesian Network are very close to what we computed directly from the dataset. The difference between results for the probability of $P(G_{CablePil}|Saturday, Hourevening)$ comes from the
first assumption, i.e., when there is no direct connection between two nodes we assume they are independent.

Note that, these conditional probabilities can also be directly calculated from the dataset but representing the connectivity between nodes using Bayesian Network makes it easier to pick features corresponding to the meaningful patterns.

Another representation example the Bayesian Network constructed from association rules is shown in Figure 4. In this network the threshold values are set such that itemsets with frequency higher than or equal to 10 have at least 50% confidence level, itemsets with frequency between 5 to 10 have 80% confidence level, and itemsets with frequency equal to 4 have 100% confidence level.

![Figure 4](image)

**TABLE IV**

<table>
<thead>
<tr>
<th>Conditional probability example</th>
<th>Result from the Bayesian Network</th>
<th>the dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(G_{Monday, Hourmorning})$</td>
<td>0.50%</td>
<td>0.50%</td>
</tr>
<tr>
<td>$P(G_{Tuesday})$</td>
<td>0.40%</td>
<td>0.35%</td>
</tr>
<tr>
<td>$P(G_{Wednesday})$</td>
<td>0.70%</td>
<td>0.60%</td>
</tr>
<tr>
<td>$P(G_{Thursday})$</td>
<td>0.80%</td>
<td>0.75%</td>
</tr>
<tr>
<td>$P(G_{Friday})$</td>
<td>0.75%</td>
<td>0.70%</td>
</tr>
<tr>
<td>$P(G_{Saturday})$</td>
<td>0.45%</td>
<td>0.50%</td>
</tr>
<tr>
<td>$P(G_{Sunday})$</td>
<td>0.55%</td>
<td>0.55%</td>
</tr>
<tr>
<td>$P(G_{Overload})$</td>
<td>0.65%</td>
<td>0.60%</td>
</tr>
<tr>
<td>$P(G_{Winter})$</td>
<td>0.70%</td>
<td>0.75%</td>
</tr>
<tr>
<td>$P(G_{Summer})$</td>
<td>0.40%</td>
<td>0.40%</td>
</tr>
<tr>
<td>$P(G_{Hourmorning})$</td>
<td>0.75%</td>
<td>0.75%</td>
</tr>
<tr>
<td>$P(G_{Hourevening})$</td>
<td>0.70%</td>
<td>0.70%</td>
</tr>
<tr>
<td>$P(G_{Hourmorning})$</td>
<td>0.75%</td>
<td>0.75%</td>
</tr>
<tr>
<td>$P(G_{Hourevening})$</td>
<td>0.70%</td>
<td>0.70%</td>
</tr>
<tr>
<td>$P(G_{Hourmorning})$</td>
<td>0.75%</td>
<td>0.75%</td>
</tr>
<tr>
<td>$P(G_{Hourevening})$</td>
<td>0.70%</td>
<td>0.70%</td>
</tr>
</tbody>
</table>

For features where both of the networks in Figure 3 and Figure 4 have the same connections, the joint probability distribution is equal. However, some joint probabilities such as $P(\text{Sunday, Hourmorning}, 2<T<3)$ is captured by direct connections in Figure 3 while the direct connection between Hourmorning and $2<T<3$ is not available in Figure 4, therefore the result will be different.

Overall, by varying the threshold values of support and confidence, we are able to control the tradeoff between accuracy of the model and its complexity.

IV. CONCLUSION AND DISCUSSION

In this paper, we give an example of the use of statistical analysis, association rules, and Bayesian Networks to explore the fault history of an electricity distribution grid in the south of Sweden. We present a simple statistical analysis of historical failures looking at the probability of occurrence and the MTBF for different failure causes and components. Then, we present analysis of relations between failure features, in particular their co-occurrences using association rules.

Finally, we show that the association rules can be used to construct a Bayesian Network that can be applied as an intuitive visualization of the conditional relations between features. In this case, most of the strong association rules can be represented by the joint probability distributions of the Bayesian Network while using only the rules with two features. Those results provide a clear and practical representation of features associated with events that can be used by managers and maintenance staff at the company.

REFERENCES


