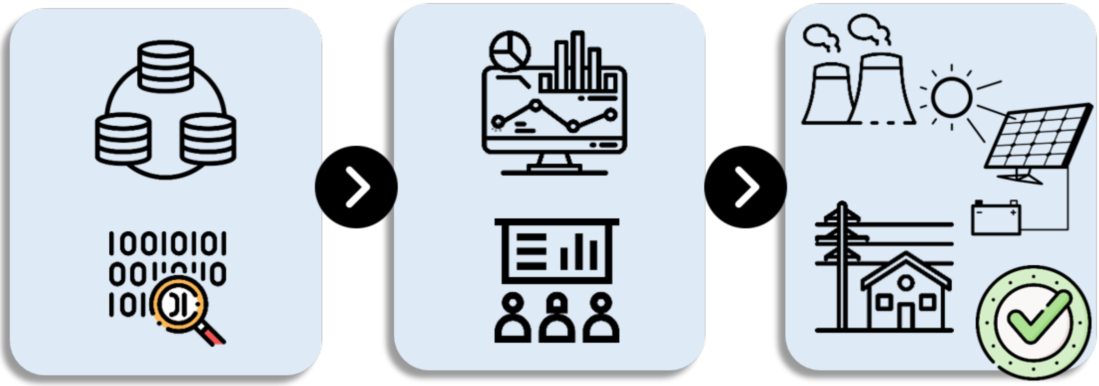
KTH ROYAL INSTITUTE
OF TECHNOLOGY

Licentiate Thesis in Electrical Engineering and Computer Science

From data collection to electric grid performance

How can data analytics support asset management decisions for an efficient transition toward smart grids?

SYLVIE KOZIEL



Stockholm, Sweden 2021

From data collection to electric grid performance

How can data analytics support asset management decisions for an efficient transition toward smart grids?

SYLVIE KOZIEL

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Licentiate of Engineering on Monday the 19th of April 2021, at 10:00 a.m. in Room Erik G. Hallén, Teknikringen 31, Stockholm.

Licentiate Thesis in Electrical Engineering and Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden 2021

Credits: Icons designed by Freepik, Smashicons, Srip, Nikita, xnimrodx, DinosoftLabs, and distributed by Flaticon www.flaticon.com

ISBN 978-91-7873-824-3
TRITA-EECS-AVL-2021:22

Printed by: Universitetsservice US-AB, Sweden 2021

Abstract

Physical asset management in the electric power sector encompasses the scheduling of the maintenance and replacement of grid components, as well as decisions about investments in new components. Data plays a crucial role in these decisions. The importance of data is increasing with the transformation of the power system and its evolution toward smart grids. This thesis deals with questions related to data management as a way to improve the performance of asset management decisions. Data management is defined as the collection, processing, and storage of data. Here, the focus is on the collection and processing of data.

First, the influence of data on the decisions related to assets is explored. In particular, the impacts of data quality on the replacement time of a generic component (a line for example) are quantified using a scenario approach, and failure modeling. In fact, decisions based on data of poor quality are most likely not optimal. In this case, faulty data related to the age of the component leads to a non-optimal scheduling of component replacement. The corresponding costs are calculated for different levels of data quality. A framework has been developed to evaluate the amount of investment needed into data quality improvement, and its profitability.

Then, the ways to use available data efficiently are investigated. Especially, the possibility to use machine learning algorithms on real-world datasets is examined. New approaches are developed to use only available data for component ranking and failure prediction, which are two important concepts often used to prioritize components and schedule maintenance and replacement.

A large part of the scientific literature assumes that the future of smart grids lies in big data collection, and in developing algorithms to process huge amounts of data. On the contrary, this work contributes to show how automatization and machine learning techniques can actually be used to reduce the need to collect huge amount of data, by using the available data more efficiently. One major challenge is the trade-offs needed between precision of modeling results, and costs of data management.

Keywords: asset management, data analytics, data management, distribution system operators, electrical power grid, machine learning, real-world datasets.

Sammanfattning

Anläggningsförvaltning inom elkraftsektorn omfattar schemaläggning av underhåll och utbyte av nätkomponenter samt beslut om investeringar i nya komponenter. Data spelar en avgörande roll i dessa beslut. Vikten av data ökar med omvandling av kraftsystemet och dess utveckling mot smarta nät. Denna licentiatuppsats behandlar frågor relaterade till datahantering som ett sätt att förbättra prestanda för anläggningsförvaltningsbeslut. Datahantering definieras som insamling, bearbetning och lagring av data. Här är fokus på insamling och bearbetning.

Först undersöks inflytandet av data på besluten relaterade till anläggningar. I synnerhet kvantifieras effekterna av datakvaliteten på utbytesstidpunkten för en generisk komponent (till exempel en ledning) med hjälp av scenariometodik och felmodellering. Faktum är att beslut baserade på data av dålig kvalitet inte är optimala. I detta fall leder felaktiga data relaterade till komponentens ålder till en icke-optimal schemaläggning av komponentutbyten. Motsvarande kostnader beräknas för olika nivåer av datakvalitet. Ett ramverk har utvecklats för att utvärdera mängden investeringar som behövs för förbättring av datakvalitet och dess lönsamhet.

Därefter undersöks sätten att använda tillgänglig data effektivt. Speciellt undersöks möjligheten att använda maskininlärningsalgoritmer på verkliga datamängder. Nya tillvägagångssätt utvecklas för att endast använda tillgänglig data för komponentrankning och felförutsägelse, vilket är två viktiga begrepp som ofta används för att prioritera komponenter och schemalägga underhåll och utbyte.

En stor del av den vetenskapliga litteraturen antar att framtiden för smarta nät ligger i stor datainsamling och i att utveckla algoritmer för att bearbeta stora mängder data. Tvärtom bidrar detta arbete till att visa hur automatisering och maskininlärningstekniker faktiskt kan användas för att minska behovet av att samla in enorma mängder data genom att använda tillgängliga data mer effektivt. En stor utmaning är avvägningarna som behövs mellan precision i modelleringsresultat och kostnader för datahantering.

Acknowledgements

This work is funded by Energimyndigheten and E.ON through the Swedish Centre for Smart Grids and Energy Storage (SweGRIDS), to which I am deeply grateful, not only for supporting my work, but also for giving me the possibility to complete inspiring research stays.

I would like to particularly thank my supervisors Docent Patrik Hilber, Dr. Per Westerlund, and Dr. Ebrahim Shayesteh for their guidance and support.

I am deeply thankful to the QED Reference Group and particularly Ola Ivarsson and Claes Ahlrot for their valuable input and stimulating meetings.

My gratitude also goes to Prof. Ryutaro Ichise for hosting me at the National Institute for Informatics, and for counseling my work.

I would like to give a heartfelt thank to Prof. Rajeev Thottappillil for reviewing my thesis and for his support; to Prof. Martin Norgren for providing a great working environment; to my colleagues at KTH with whom I have such pleasure working; to Brigitt Högberg, Eleni Nylén, Carin Norberg and people whithin the IT support, the service center, and KTH library, for patiently helping me with all the collateral tasks and various issues I encounter.

My sincere thanks go to the members of Ichise laboratory at the National Institute for Informatics (NII) in Tokyo for the enlightening talks and kind welcome. Thank you for making it such a lovely stay.

I am hugely indebted to my previous colleagues at Indicta and Prognos AG, as well as to my previous teachers at Lycée Henri Wallon, Lycée Saint-Louis, AgroParisTech, Universitat Politècnica de València, IFP Énergies nouvelles, and Institut national des sciences et techniques nucléaires (INSTN), whose dedication and passion contributed to develop my thirst for knowledge and curiosity for differences and the unknown.

Acronyms

List of commonly used acronyms:

AE	auto-encoder
AUC	area under the curve
CONV	convolutional
ENS	energy not supplied
LSO	local system operator
MLP	multi-layer perceptron
OH	overhead
RF	random forest
SMOTE	synthetic minority oversampling technique
SVM	support vector machines
UG	underground
USD	United States dollar
XGB	extreme gradient boosting

List of Papers

I ***Investments in data quality: Evaluating impacts of faulty data on asset management in power systems***
Sylvie Koziel, Patrik Hilber, Per Westerlund, Ebrahim Shayesteh
Applied Energy, volume 281, pp. 116057 (2021)

II ***Application of big data analytics to support power networks and their transition towards smart grid***
Sylvie Koziel, Patrik Hilber, Ryutaro Ichise
IEEE International Conference on Big Data (Big Data), pp. 6104-6106 (2019)

III ***A review of data-driven and probabilistic algorithms for detection purposes in local power systems***
Sylvie Koziel, Patrik Hilber, Ryutaro Ichise
International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), pp. 1-6 (2020)

IV ***Component ranking and importance indices in the distribution system***
Sindhu Kanya Nalini Ramakrishna, Sylvie Koziel, Patrik Hilber, David Karlsson, Gustav Stenhag
Accepted in PowerTech Conference 2021

V ***Detecting rare events for low frequency, sequential and unspecific datasets: application to failure prediction of an HVDC line***
Sylvie Koziel, Patrik Hilber, Ryutaro Ichise
Submitted to IEEE Transactions on Systems, Man, and Cybernetics

Other contributions by the author not included in the thesis.

VI ***Forecasting cross-border power exchanges through an HVDC line using dynamic modelling***
Sylvie Koziel, Patrik Hilber, Per Westerlund, Ebrahim Shayesteh
IEEE International Conference on Big Data (Big Data), pp. 4390-4394 (2019)

I am the main author of **Papers I**, **II**, **III**, and **V**. The idea of **Paper I** came from my supervisor P. Hilber. I performed the conceptualization, methodology, investigations, coding, visualization, and writing. My supervisors P. Hilber, E. Shayesteh and P. Westerlund supervised the work, and contributed to the methodology, the interpretation of results and reviewing the paper. The idea and methodology of **Papers II, III, IV and V** came from myself. I performed the literature review and writing of **Papers II, III, V**, under the supervision of P. Hilber and R. Ichise. The idea of **Paper IV** came from discussions with O. Ivarsson (E.ON) and C. Ahlrot (E.ON). **Paper IV** was written by Sindhu Kanya Nalini Ramakrishna, and is the result of a master thesis that was carried out under my supervision, and in collaboration with Digpro. She implemented the methodology and realized the literature review.

The subsection ”Substation-level failure detection” of section 5.2 is based on another part of the master thesis of Sindhu Kanya Nalini Ramakrishna, cited as reference [1], and is the result of a collaboration between the student and myself. The workdiva has been presented at the student's master thesis defense.

Contents

Chapter 1

Introduction

1.1 Background and motivation

Traditional power systems are composed of a few centralized large power plants providing most of the electricity to the consumers. However this structure, defined by unidirectional power flows from big power generators to consumers, is undergoing changes. The need to decarbonize energy to fight climate change has supported the development of climate-friendly ways to generate electricity (wind turbines, solar panels), to transport people and goods (electric vehicles), to heat spaces (heat pumps). The deployment of distributed generation, the apparition of new loads, and the multiplication of prosumers, not only change the structure of the grid, but also the way power adequacy and security of supply are calculated and managed [2–5]. The power system integrates an increasing number of new devices, components and stakeholders, becomes more complex, and is much more difficult to analyze.

At the same time, the role of data in the power system is growing. For instance, an increasing number of sensors and smart meters are installed and provide more data [6–8]. This context creates a situation where the use of data for asset management is both possible and needed. Decisions increasingly need to be supported with data-driven approaches. In this transition toward ”smart grids”, data analytics play a central role. Therefore, the study of the ways data collection and processing can support and improve asset management decisions is primordial to decarbonize the power system, while keeping costs and the security of supply at satisfactory levels.

1.2 Research objectives

Literature related to big data analytics is well developed, and expanding. Still, the data used often do not reflect the actual data available to public and private organizations, which are often incomplete, faulty, sparse, imbalanced and unspecific. Some literature on data analytics also use data that are not collected by all asset operators, or are supposed to be available in the future.

This work focuses on commonly available data, and on approaches to manage data in the most efficient way, to improve decision making in asset management, and ultimately support the transition to smart grids. Data management includes data collection, data processing (data analytics), and data storage. One major challenge is to determine the data quality, type and quantity needed for the assets to achieve a given performance level.

1.3 Research contributions

In the framework of this thesis, the relations between data management and grid performance have been conceptualized (Figure 1.1), and are explained in Chapter 2. *Data* and information on the *situation* of the grid are inputs used in *approaches*. *Data* refer to physical measurements on the grid such as load, power flows, and outages, while *situation* refers to regulatory requirements, as well as the structure of power generation (e.g. share of renewables) and consumption (e.g. number of electric cars). *Approaches* include data processing and analytics. Data and information are processed in *approaches*, which provides outputs that are used to take *actions* related to asset management. The impacts of *decisions* can be measured through *performance* indices.

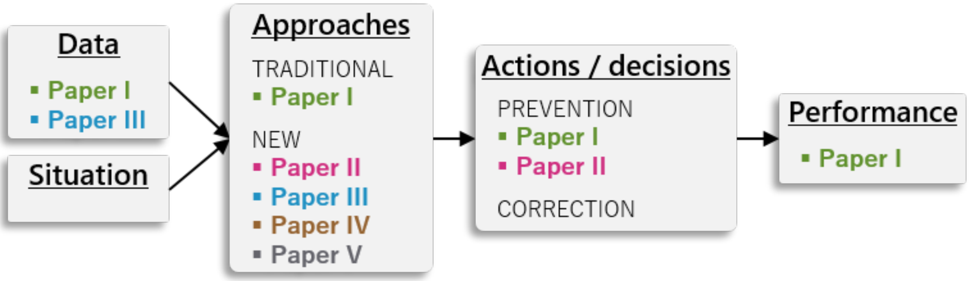


Figure 1.1: Research framework and scope of papers within the framework.

- The impacts of *data* quality on decisions related to component replacement, and then on grid *performance* (in terms of costs) are studied in **Paper I**. We propose an innovative data quality management framework enabling asset managers: (i) to quantify the impact of poor data quality, and (ii) to determine the conditions under which an investment in data quality improvement is required.
- The relationships between *data*, *approaches* and *actions*, are identified in **Papers II** and **III**. We discuss where and how machine learning *approaches*

could contribute to more efficient asset management *decisions* (**Paper II**). We also provide a literature review focused on three particular *approaches*: anomaly detection, fault location, and load disaggregation, and analyze them in terms of *data* requirements (**Paper III**).

- *New approaches* supporting asset management decisions are developed in **Papers IV** and **V**. We present methods to rank grid components according to their importance, using outage data. Importance indices enable to prioritize components according to a chosen criterion, and to adapt monitoring strategies (**Paper IV**). We develop a component failure prediction model without component-specific sensor data (**Paper V**), and study failure detection at substation level.

1.4 Research ethics

This thesis intends to support the transition toward smart grids. In fact, it contributes to helping grid managers to take efficient asset management decisions, especially taking into account the development of renewable energy sources and electric cars. It also aims to reduce the need for data collection and processing to the minimum required to reach a satisfactory performance of the grid, thus limiting the pollution generated by big data collection and storage, as well as heavy computations. In this way, this work fulfills my moral responsibilities toward the society and the environment. Finally, the data used in this work come from publicly available sources (in which case they are clearly stated), or from private utility data (in which case, they are not displayed individually), so that transparency and confidentiality are guaranteed.

1.5 Thesis organization

The rest of the thesis is organized as follows:

- Chapter 2 provides an analysis of the role of data in asset management in power systems, and explains in detail Figure 1.1 (**Paper I**).
- Chapter 3 demonstrates how changes in data quality affect grid performance, and illustrates the relations between *data*, *approaches*, *actions* and *performance* in Figure 1.1 (**Paper I**).
- Chapter 4 explains in which ways the energy transition affects data and grid management (elements in Figure 1.1), and how machine learning can support the transition (**Papers II, III**).
- Chapter 5 describes new *approaches* that have been developed to support asset management decisions (**Papers IV and V**).
- Chapter 6 contributes to the discussion and concludes the thesis.

Chapter 2

Relations between data and grid performance

This chapter explains how data are used in practice to take decisions related to assets. The aim is to show the central role played by data, especially to reach a given performance level for the grid. The chapter concludes by displaying the research framework.

2.1 Data as the basis to achieve efficient asset management decisions

Performance goals

The main task of power grid operators is to manage their assets in such a way that they achieve a given level of grid *performance*. The assets are a set of components that compose the electric power grid. Grid performance can be defined in several ways, but it generally encompasses three aspects:

- A technical aspect: i) Grid operators should prevent outages to minimize the frequency of power supply interruptions; ii) if outages happen, their duration as well as impacts or severity should be reduced to the minimum. Here, the customer importance should be taken into account; iii) power quality, which includes flickers, harmonic distortions, voltage and frequency instabilities, should be kept at an acceptable level.
- An economic aspect: Asset management costs, including investments, maintenance and reparations, should be minimized.
- A regulatory aspect: Regulations related to safety, environmental, technical and economic requirements (for example amortization time allowed, interruption fines, type of allowed investments) should be fulfilled. This aspect put constraints on the two aspects mentioned above.

Grid performance can be measured through a variety of indices such as: the number of outages, the duration of outages, the energy not supplied, the costs of maintenance, or voltage instabilities.

Asset management decisions (actions)

Asset management consists in taking a number of *actions* or *decisions* that belong to one of the following categories:

- *Preventive* actions: They aim at avoiding outages from happening, and entail the scheduling of components maintenance and replacement, as well as the investment in new components (for example new lines, new transformers, dynamic line rating, switch placement, or automatization);
- *Corrective* actions: They aim at reducing the duration and impacts of outages when these could not be avoided. They include fault localization, network reconfiguration, supply curtailment, and peak shaving.

Approaches

The *decisions* are taken based on the insights provided by *approaches*. Traditionally, approaches for asset management decisions have been data-free or time-based. This means that assets are maintained or replaced after a fixed period of time, and not based on their health condition. On the contrary, data-driven approaches are based on a monitoring step, where data and information are collected, and a modeling step, where the collected information is processed. Approaches are specific to the decision they are supporting, as illustrated below.

In the case of *investment* decisions, one task might be to calculate the supply adequacy for the future. The approach consists in modeling consumption and supply patterns, and generating forecasts to calculate the adequacy of supply [9]. The input data include load time series and power flows. Another task might be to choose between investment options. The approach then consists in modeling the system, and running an optimization algorithm that gives the optimal investment option corresponding to the minimum costs [10]. The input data include costs, outage data, and power flows.

In the case of *maintenance/replacement scheduling*, a common approach is to assess the risk related to the failure of a particular component, as explained in **Paper I**. This is often evaluated by two variables:

- *the probability of failure*, which gives information on the condition of the component (state of deterioration). Various methods can evaluate the condition: scoring systems [11], semi-Markov modelling [12], distribution functions [13] [14]. Failure distributions are the most common way to model the condition of the component;

- *the consequence of the component's failure on the system*, which takes into account the function and relative importance of the component in the system compared to others (criticality). Various methods are used to determine the importance level: scoring systems [11], fixed costs of failure and replacement [15], and criticality importance indices [16], [17]. An index can be calculated for example through a sensitivity analysis of system reliability (measured by the energy not supplied or ENS) to component reliability (measured by the unavailability due to failure).

Combining the probability of failure and the importance index enables power grid managers to classify and prioritize their components, and thus supports the scheduling of maintenance and replacement. The input data may consist in sensor data, inspection results, and outage statistics.

2.2 Research framework

Figure 2.1 summarizes the relations between the concepts detailed in Section 2.1. It represents a value chain: raw *data* and *situational information* are used in *approaches* and transformed into information with higher value. This produced information is then used to take asset management *actions*, which should enable to achieve grid *performance*. This value chain represents the research framework. Each link of the chain is illustrated in the figure by a non-exhaustive list of major elements.

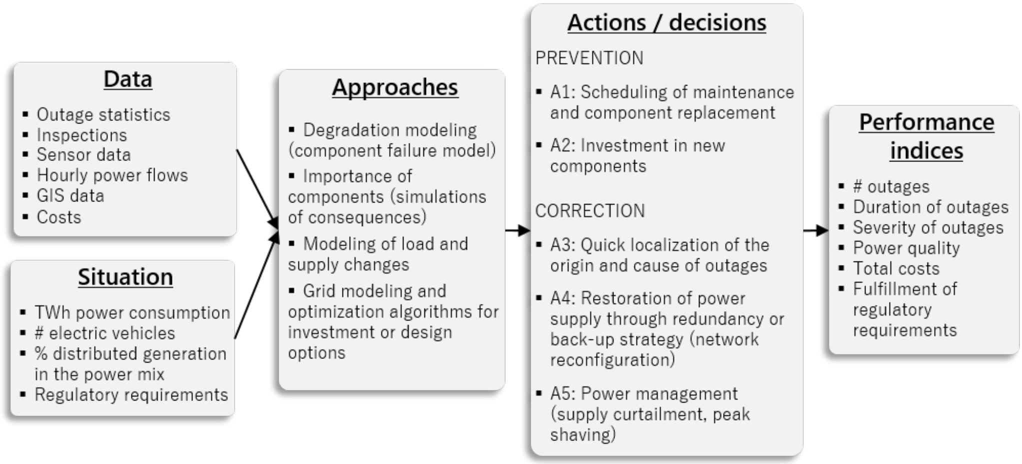

Data
▪ Outage statistics
▪ Inspections
▪ Sensor data
▪ Hourly power flows
▪ GIS data
▪ Costs
Situation
▪ TWh power consumption
▪ # electric vehicles
▪ % distributed generation in the power mix
▪ Regulatory requirements
Approaches
▪ Degradation modeling (component failure model)
▪ Importance of components (simulations of consequences)
▪ Modeling of load and supply changes
▪ Grid modeling and optimization algorithms for investment or design options
Actions / decisions
PREVENTION
▪ A1: Scheduling of maintenance and component replacement
▪ A2: Investment in new components
CORRECTION
▪ A3: Quick localization of the origin and cause of outages
▪ A4: Restoration of power supply through redundancy or back-up strategy (network reconfiguration)
▪ A5: Power management (supply curtailment, peak shaving)
Performance indices
▪ # outages
▪ Duration of outages
▪ Severity of outages
▪ Power quality
▪ Total costs
▪ Fulfillment of regulatory requirements

Figure 2.1: Conceptualized relations between data collection and grid performance.

The research framework constitutes the guiding thread for this work, and is used to study the relations between data and grid performance. In particular,

- The way a change in *data* quality impacts *decisions* (asset management actions), and *performance* (in terms of costs) is explored using usual *approaches* - Chapter 3.
- The influence of the energy transition on the framework is investigated. Especially, big *data* analytics offer new possibilities to process data, and thus could be used to improve asset management *decisions* - Chapter 4.
- New *approaches* to improve the efficiency of *decision* making by giving more insights into the state of the grid are developed - Chapter 5.

Chapter 3

Impacts of data quality on asset management decisions

This chapter is based on **Paper I**, and focuses on the component replacement time as asset management decision. Data is at the core of data-driven replacement decisions. In practice, the quality of data varies from very good to severely lacking. For example, it can be incomplete, inaccurate, incorrect or missing. Therefore, the quality of data may have a significant influence on the efficiency of asset management decisions. This work contributes to the reflection on the value of data, and provides a method to quantify the impact of poor data quality on asset management decisions.

3.1 Optimization of component replacement time and analysis of key influencing factors

Theoretical model

The model used to find the optimal year of replacement is based on i) a failure distribution that models the condition of the component, and ii) fixed costs of failure and replacement to take into account the impact of outages on the system (see traditional approaches in section 2.1). One novelty in the proposed optimization method is that it integrates discount rates that take into account the time value of money.

To model the risk that component i fails before its planned replacement in year t, the cumulative function F of the Weibull distribution is used, which is commonly employed to model assets wear-out. The probabilities of failure are conditional to the age of components, to take into account the fact that components have not failed before the beginning of the planning period. This is done by dividing the expressions in (3.1) and (3.2) by the same probability which cancels out in (3.3). Therefore, only the simplified expressions are presented here.

The probability that the component i fails before time t is modelled by:

$$P_i(t) = F_i(t; \alpha, \beta) = 1 - \exp\left(-\left(\frac{t + a_i}{\beta}\right)^{\alpha}\right) \tag{3.1}$$

where:
t is the planned year of replacement
a_i is the age of the component i
α denotes the shape parameter of the Weibull distribution
β denotes the scale parameter of the Weibull distribution
α and β are constant scalars and assumed to be known. α is set, and β is calculated based on the value of α and the average technical lifetime of the components.

The probability for the component i to fail exactly at year k is modelled by:

$$\begin{aligned} p_i(k) &= F_i(k; \alpha, \beta) - F_i(k-1; \alpha, \beta) \\ &= \exp\left(-\left(\frac{k - 1 + a_i}{\beta}\right)^{\alpha}\right) - \exp\left(-\left(\frac{k + a_i}{\beta}\right)^{\alpha}\right) \end{aligned} \tag{3.2}$$

The optimal asset management decision for component i corresponds to the year of replacement T_i with the lowest annual costs over the whole period:

$$T_i = \arg\min_t \frac{\frac{C_r}{(1+r)^t} \times (1 - P_i(t)) + \sum_{k=1}^{t} \frac{C_r + C_i}{(1+r)^k} \times p_i(k)}{(t + a_i) \times (1 - P_i(t)) + \sum_{k=1}^{t} (k + a_i) \times p_i(k)} \tag{3.3}$$

where:
C_r denotes the costs of replacing the component at the planned year
C_i denotes the additional costs generated by a failure of the component before the planned year of replacement, including unplanned interruption of supply
r is the discount rate

The numerator represents the average costs. The first term accounts for the costs of replacing a component at the planned year, weighted with the probability that the component does not fail before replacement. The second term represents the costs incurred if the component fails before the year of replacement. The denominator represents the average lifetime of the component.

Analysis of influencing factors

The optimal year of replacement depends on the chosen parameters, particularly the discount rate r, and the ratio between unplanned (corrective) interruption costs and planned (preventive) replacement costs $(C_r + C_i)/C_r$:

- When the discount rate increases, the effect of high interruption costs is reduced, and therefore the optimal year of replacement is postponed. When the discount rate reaches a breaking point, the optimal decision in economic terms is to let the component run to failure;

- The higher the corrective costs compared to the preventive costs, the sooner the component should be replaced. When the unplanned failure of the component does not generate any additional expenses (ratio equal to one), the optimal management decision is to let the component run to failure.

3.2 Quantification of the impact of poor data quality on component replacement time

Theoretical formulation of the problem

Data quality is defined in [18] as ’fitness for use’, meaning that it is a relative assessment of the extent to which data serve the purpose of the user. Thus, the level of data quality can be sufficient for a given task, but not good enough if the task changes or the purpose evolves. This also means that the level of data quality should be monitored over time to control its adequacy to the tasks.

To quantify the costs of faulty data, three scenarios are developed. In the *Reference* scenario, the real condition of components, approximated by their age, is known and used to calculate the optimal replacement time. In the *Imperfect information* scenario, the assessed condition is only partially correct, which means that the data quality is lower. Some components have been assessed as being in a worse condition than they are in reality. Two variables describe the level of data quality: the share of components affected by incorrect data (‘share of faulty data’), and a variable determining, for each affected component, the difference between its real age and its assessed condition, or in other words, how far the incorrect data diverge from the true value (‘deviation from the true value’).

Results of simulations

Because faulty data have been translated in this work by an overestimation of the age of the component, poor data quality affects the asset management decision by hastening the planned year of replacement by one or more years. The interesting result of the simulation is the quantification of the impact of faulty data. In particular, this method provides a useful assessment of the orders of magnitude of such time shifts, as a function of components’ age and data quality (expressed by the ‘share of faulty data’ and ‘deviation from the true value’).

When considering a population with uniformly distributed age categories between 0 and 30 years, poor data quality can shift the planned year of replacement by 0.03 years (5 % of components with incorrect data deviating by 5 % from the true value) to 2.2 years (50 % of components with faulty data deviating by 140 % from the true value).

3.3 Decisions of investment in data quality improvement

Theoretical formulation of the problem

Poor data quality generates costs, but so does improving data quality: prevention costs (training, monitoring, deployment), detection costs (sensors, analysis, reporting), and repair costs (planning and implementation). A major challenge for grid managers is to find a balance between these costs. This is equivalent to identify the optimal level of data quality. This section aims to provide a decision support tool that helps grid managers estimate the amount of investments in data quality improvement that would be profitable.

To this end, different levels of investments are introduced in an *Investment in higher data quality* scenario, and their effect on the year of replacement is analyzed. The investment leads to a more accurate prediction of the time to replace components compared to the *Imperfect information* scenario. For grid managers to invest in data quality improvement, the savings generated by investments must overcompensate the initial investment (first term of equation (3.4)) and yearly costs of improved data quality (second term of equation (3.4)). The savings are the avoided costs of the asset management without investments (third term of equation (3.4)).

$$
\begin{aligned}
Benefits = &-(p \times C_r) - \left((T - \widetilde{t_f}) \times \frac{C_r}{(1+r)^{\widetilde{t_f}}} \times \frac{1}{\widetilde{t_f} + a} \right) \\
&+ \left((T - t_f) \times \frac{C_r}{(1+r)^{t_f}} \times \frac{1}{t_f + a} \right)
\end{aligned}
\tag{3.4}
$$

where:
a is the age of the component
r is the discount rate
p is the percentage of component's replacement costs
C_r are the costs of replacement of the component
T is the planned year of replacement in the *Reference* scenario
t_f is the planned year of replacement in the *Imperfect information* scenario
$\widetilde{t_f}$ is the planned year of replacement in the *Investment in higher data quality* scenario

The second term represents the discounted asset management costs in the *Investment in higher data quality* scenario. They are expressed as the reduction of the component's lifetime due to partly incorrect data, multiplied by the yearly revenues generated by the component. The third term represents the avoided and discounted costs of asset management in the *Imperfect information* scenario. In this configuration, the investment is profitable when the sum of avoided costs exceeds initial investment and yearly costs of data quality improvement.

Results of simulations

Figure 3.1 illustrates the kind of results that can be obtained when using the framework. Up to a certain level of data quality, economic gains are negative, meaning that investments in higher data quality do not offset economic gains resulting from improved asset replacement decisions. Data quality has to be 'sufficiently poor' for the investment to be profitable.

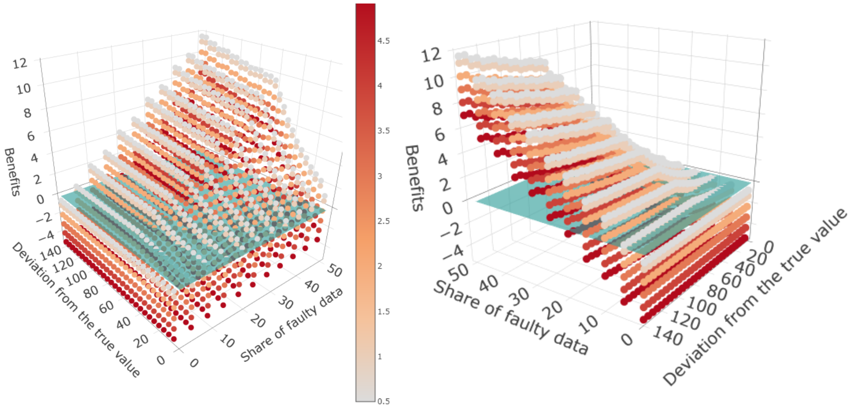


Figure 3.1: Determination of thresholds for profitable investments in improved data quality. Left and right panels represent the same plot seen from different perspectives. Investments are represented by a color scale, and given as a percentage of the component's costs. The deviation from the true value and share of faulty data are expressed as percentages on the axes. Benefits are given in 1000 USD.

3.4 Implementation of the framework in practice

Data quality can be improved by measures that enable a better evaluation of the condition of components. These measures are diverse, and include the development of a data quality management *strategy* on the one hand, and concrete *actions* such as investigations or the installation of measurement devices on the other hand. Figure 3.2 illustrates these concepts.

The framework has been implemented in one example. Assuming a population of overhead lines, the asset manager aims to optimize the replacement time of the lines based on the available data. Phase 1 of the framework reveals that around 20 % of the data related to the age of the lines are faulty, and that the age has been overestimated by 50 % on average. The asset manager wants to know if investing in a device that would improve data quality would be economically profitable. Using the results of phase 1 and the model, the impact of faulty data on the replacement year is assessed (phase 2). The asset manager then plans to invest in a device that measures dissipation/power factor and capacitance, which indicates

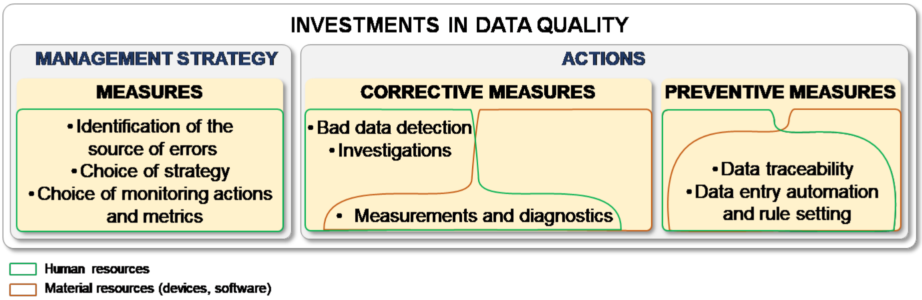

INVESTMENTS IN DATA QUALITY
MANAGEMENT STRATEGY
MEASURES
• Identification of the source of errors
• Choice of strategy
• Choice of monitoring actions and metrics
ACTIONS
CORRECTIVE MEASURES
• Bad data detection
• Investigations
• Measurements and diagnostics
PREVENTIVE MEASURES
• Data traceability
• Data entry automation and rule setting
Human resources
Material resources (devices, software)

Figure 3.2: Overview of possible measures to improve data quality.

the overall condition of the insulation of the lines. This estimated condition can be translated into an equivalent age.

Three artificially generated scenarios with different assumptions related to population size, line length, share of faulty data, deviation from true value, cost and type of monitoring device are developed. They aim to illustrate how the benefits of an investment into data quality improvement change in different situations (Table 3.1). Many other scenarios could be generated in this way.

Table 3.1: Framework implemented in a practical case for different scenarios.

Variable	**Scenario 1**	**Scenario 2**	**Scenario 3**
Number of power lines	100	10	10
Average length of lines in km	1	2	2
Share of faulty data in %	20	10	10
Deviation from true value in %	50	20	20
Costs of condition monitoring device in thousand USD	70	70	70
Condition monitoring type (T = transportable)	T	T	non T
Profits/losses from investing in data quality improvement in thousand USD per line	**1.11**	**-0.13**	**-0.73**

Only scenario 1 corresponds to a situation where the investment in a device would be profitable. Having a smaller population size and high data quality (low share of faulty data and deviation from the true value) reduces the benefits of investing in data quality, as illustrated in the differences between scenarios 1 and 2. Transportable devices can be used on several components and lower the costs of investments, explaining the difference between scenarios 2 and 3.

Chapter 4

Increasing importance of data due to the energy transition

The aim of this chapter is to explain how the transition toward cleaner energy sources impacts the power grid, as well as the value chain illustrated in Figure 2.1, and how data-driven approaches can both contribute to the transition and support performance goals. This chapter is based on literature reviews carried out in **Papers II** and **III**. Some of the identified new approaches are implemented in Chapter 5.

4.1 Complexification of the traditional asset and data management

The decarbonization of the energy sector is at the heart of the fight against climate change [19]. This energy transition translates into the replacement of large fossil fuel power plants by smaller-scale renewable energy-based power generators such as wind turbines, solar panels, and hydropower, the replacement of fossil fuel vehicles by electric or fuel cell vehicles, and the replacement of fossil-based heating by, among others, heat pumps or biomass heating. This transition heavily impacts the power grid: intermittent and distributed renewable generation such as wind farms and solar panels are changing power supply patterns, and the development of electric vehicles are not only influencing loads, but also transforming unidirectional power flows into bi-directional ones [20] [21]. The *situation* of the grid, represented in Figure 2.1, changes.

As the structure of the power grid complexifies, integrating more and more loads and generators, it becomes increasingly difficult to operate and manage it, using the traditional methods. The energy transition generates challenges for asset and data management:

- Traditional *data* might not properly account for the presence of new devices and stakeholders, and their impacts, for example on load profiles, supply forecast, or power flows. More data need to be collected. One challenge is to select the type and amount of data that are needed.

- Traditional *approaches* are not adapted to rapid changes affecting the grid. New processing methods are needed to take into account the new data streams, to support decisions in a pro-active way, and enable automated decisions.

- Asset management *actions* such as preventive (scheduling maintenance and replacement, investments) and corrective actions become challenging. In fact, more variable power flows put a new strain on grid components, possibly changing maintenance scheduling; rapidly changing load patterns complexify the calculations of power supply adequacy, and therefore investment decisions; voltage and frequency instabilities induced by high penetration of renewables make the adjustment of corrective actions and balancing measures necessary.

4.2 New approaches offered by big data analytics to support asset management decisions and the transition to smart grids

The development of big data analytics, and especially machine learning has become of high interest to many countries and companies. The integration of machine learning to support efficient and automated asset management decisions is part of the transition to smart grids. Grid managers could take advantage of advances in information and communication technologies (such as smart meters, sensors, 5G, processors), and harness consumer data, weather data, data from renewable power generators, and from other internet sources, to tackle the challenges of fluctuating power generation and loads.

While research is quite extensive in real-time power operations and pricing, as outlined in **Paper II**, few researchers have applied machine learning to help distribution companies manage their assets, and adapt them to the greener electricity generation and consumption patterns on a more strategic timeline. In **Paper II**, the possible ways machine learning could be applied in power systems to improve asset management strategies are presented. Three main areas have been identified in the asset management process, where these techniques can support decisions: detection, prediction, and selection.

Detection of changes in patterns or anomalies

This detection function is fundamental for asset managers, who need to identify early signs of changes well in advance, to adapt the network and plan investments. Especially, some factors like increased population, installation of renewables and deployment of electric vehicles can increase instabilities and outage risks if they are not detected early enough. Therefore, an algorithm that would detect any changes to usual or normal operating modes, and identify the origin or cause of these changes is required for the reliable operation of the future grid. This kind of task can be seen as a classification or novelty/bad data detection problem.

Improved predictions

In the field of asset management, and in particular maintenance, useful predictions are the predictions concerning component failures. The elaboration of a model representing the component's failure requires a deep understanding of failure causes, and component degradation according to its operational characteristics and environmental conditions. Often, the data needed to build the failure rate model are not available, or difficult to obtain. Therefore, some research activities are focusing on algorithms to detect signals that would help predict outages, on the basis of commonly available data (for example history of failures, data on maintenance activities, power flow measurements, or weather-related data). This approach is developed in Chapter 5.2.

Selecting efficient asset management options

Usually, the selection of possible options is made based on an optimization algorithm. However, the power system is expected to become more complex because of bidirectional flows, as well as the development of distributed micro-generation. Therefore, usual (linear, non-linear, mixed-integer) programming methods used to model power flows are likely to get exceedingly complex for the network manager to implement. Machine learning could offer a new way of analyzing for example the benefits or drawbacks of "smart" technologies that allow flexibility (such as flexible alternating current transmission systems, demand side management, or quadrature boosters) compared to traditional adjustments, which are often capital-intensive and not flexible (such as network expansion, and grid reinforcement). This would help avoiding stranded assets, which are common in environments that are characterized by high uncertainty and changing rate. One example of a selection method, applied to component ranking, is given in Chapter 5.1.

4.3 Algorithms for detection purposes adapted to real-world datasets

Often, machine learning algorithms suppose the availability of component-specific sensor data. In practice, grid operators have access only to aggregated or partial signals, from which the relevant information is difficult to extract. Also, installing sensors at all components in the distribution systems is unrealistic. At the same time, distribution system operators are mostly impacted by the deployment of distributed generation. Besides, many algorithms require data collected at a high sampling rate. However, grid operators usually have hourly or half-hourly data, and do not measure or store data with higher sampling rate. Therefore, algorithms from the field of big data analytics should be adapted to real-world data, characterized by low specificity (aggregated/partial signals), and low sampling rate.

Paper III provides an analysis of detection algorithms developed in power distribution systems for real-world data. Signal detection is a part of data analytics that "deals with the processing of information-bearing signals for the purpose of extracting information from them" [22, p.1]. The analysis is focused on detection fields that are relevant for power systems, namely anomaly detection, fault location, and load disaggregation. The algorithms are classified according to their type. The way they are implemented is analyzed. Especially, we aim to clarify which types of detection algorithm can be used for which task.

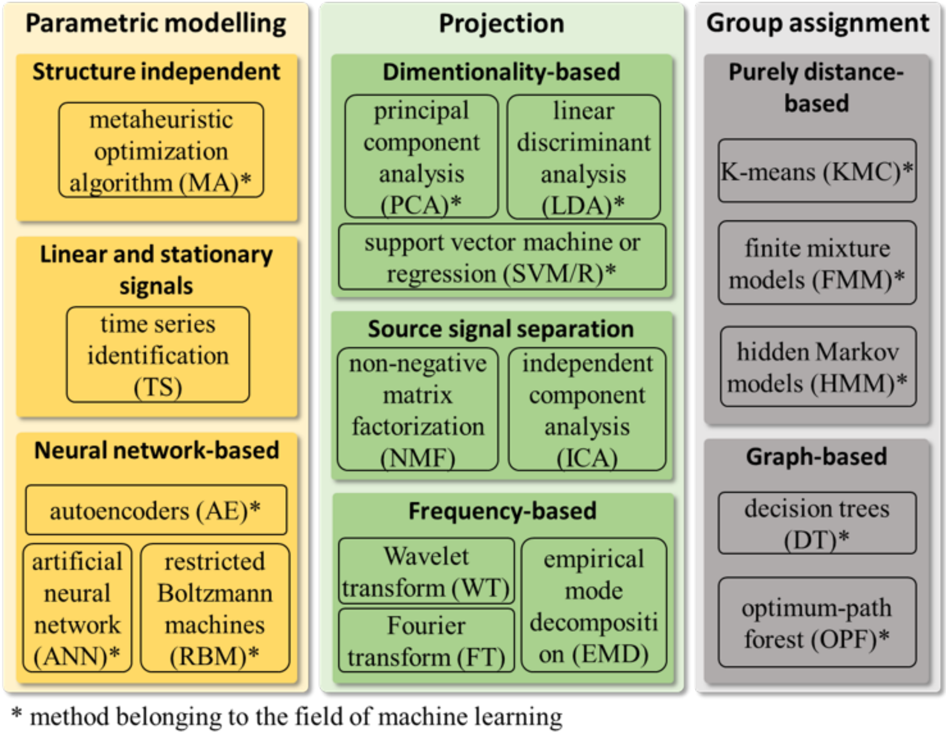
Classification of detection algorithms

The algorithms are classified in three types, as shown in Figure 4.1.

Algorithms within the "parametric modelling" type model the relationships between variables through function approximation. They use different structures or function characteristics: i) linear and stationary signals (time series), ii) neural networks (autoencoders, other neural networks, restricted Boltzmann machines) or iii) structure-independent optimization algorithms.

Algorithms within the "projection" type project the data into other dimensions such that features can be extracted and data separated more easily. Various transformation processes are used: i) dimensionality reduction using eigenvectors (linear discriminant analysis, principal component analysis), ii) dimensionality increase to better separate data (support vector machines), iii) separation of a set of source signals from a set of mixed signals (non-negative matrix factorization, independent component analysis), or iv) projection of data from the time to the frequency domain (Fourier / wavelet transform and empirical mode decomposition).

Algorithms within the "group assignment" type assign a cluster to each observation such that the formed clusters are homogeneous. Two methods can be used: i) distance-based partitioning, which minimizes the sum of squared distance between centroids and observations through the expectation-maximization algorithm (K-means, finite mixture models, hidden Markov models), and ii) graph-based partitioning, which uses tree structures in addition to metrics such as Gini impurity, information gain, variance reduction (decision trees, random forests, optimal path forest).



Parametric modelling
Structure independent
metaheuristic optimization algorithm (MA)*
Linear and stationary signals
time series identification (TS)
Neural network-based
autoencoders (AE)*
artificial neural network (ANN)*
restricted Boltzmann machines (RBM)*
Projection
Dimentionality-based
principal component analysis (PCA)*
linear discriminant analysis (LDA)*
support vector machine or regression (SVM/R)*
Source signal separation
non-negative matrix factorization (NMF)
independent component analysis (ICA)
Frequency-based
Wavelet transform (WT)
Fourier transform (FT)
empirical mode decompositi on (EMD)
Group assignment
Purely distance-based
K-means (KMC)*
finite mixture models (FMM)*
hidden Markov models (HMM)*
Graph-based
decision trees (DT)*
optimum-path forest (OPF)*
* method belonging to the field of machine learning

Figure 4.1: Detection methods for anomaly detection, fault location, and load disaggregation, classified according to their type.

Implementation of algorithms for anomaly detection, appliance-specific load detection, and fault location

Figure 4.2 shows the main tasks that are performed when implementing the algorithms of the three detection fields. Green boxes represent the input, blue boxes the data processing part, and pink boxes the results.

Some algorithms are ubiquitous. They perform multiple tasks, and are seldom used in combination with other techniques: metaheuristic algorithms, autoencoders / restricted Boltzmann machines, optimal path forests. Other algorithms are often combined, and used to fulfill a specific task, as indicated in Table 4.1.

Some typical challenges include determining the threshold to label data as anomalous, dealing with noisy, incomplete and/or imbalanced data, and analyzing online streaming data (usually dealt with using sliding windows or parallel processing).

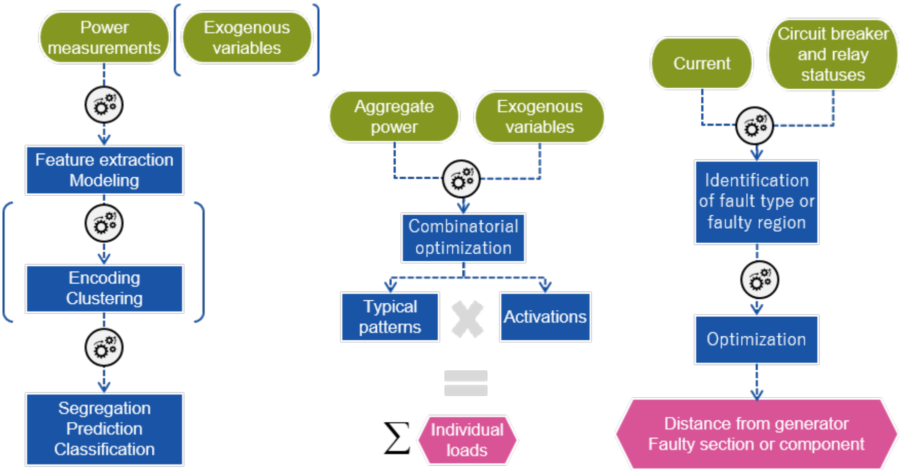

Power measurements
Exogenous variables
Feature extraction Modeling
Encoding Clustering
Segregation Prediction Classification
Aggregate power
Exogenous variables
Combinatorial optimization
Typical patterns
Activations
Individual loads
Current
Circuit breaker and relay statuses
Identification of fault type or faulty region
Optimization
Distance from generator Faulty section or component

Figure 4.2: Application of detection algorithms in power systems. From left to right: anomaly and theft detection, appliance-specific load detection, fault location. The step represented between brackets is implemented only in certain cases.

Table 4.1: Algorithms performing specific tasks.

Algorithms	**Tasks often performed**
Time series, Fourier transform, wavelet transform	Upstream tasks like feature extraction
Support vector machines	Classification (anomaly detection)
Principal component analysis, linear discriminant analysis, K-means, decision trees	Both feature extraction and classification
Neural networks, independent component analysis, matrix factorization, hidden Markov models, finite mixture models	Downstream tasks in signal disaggregation and fault location

Chapter 5

Development of new approaches to improve decision making

This chapter provides reflections on the way to use available data to create relevant new information for grid operators. The focus of this chapter is to develop valuable insights related to individual components, without having to install component-specific meters and sensors. The reason is that, for a system composed of a great number of components (like at the distribution level), the costs of monitoring each of the components would probably be higher than the benefits of the monitoring resulting from efficiency gains of asset management decisions. This trade-off between the costs and the benefits of data management is at the heart of the present reflection.

This chapter focuses on two of the new approaches identified in Chapter 4.2: selection (of critical components according to their importance), and prediction (of component failure).

5.1 Selection of critical components through component importance indices using outage data

This section is based on **Paper IV** and [1].

Importance indices

Monitoring the condition of components helps taking preventive actions to avoid failures, and increases reliability. However, performing such monitoring for all components of the distribution grid is prohibitively expensive. Instead, distribution system operators could focus efforts only on the most critical components. In particular, importance indices enable to prioritize components according to a chosen criterion, and to adapt monitoring strategies. Existing methods to calculate component importance index are discussed in **Paper IV** and [16].

In this work, two types of empirical methods for ranking components are developed, as represented in Figure 5.1:

- the method based on the *calculation of de-energization time*, which takes into account only switch events recorded during the outages, and does not take into consideration the component that caused the outage;
- the methods based on the *identification and localization of components that are responsible for outages*. Four methods belong to this type, which rank the responsible components according to their failure frequency or the impact that their failure has on the system (disconnected power, energy not supplied (ENS) and customer outage time).

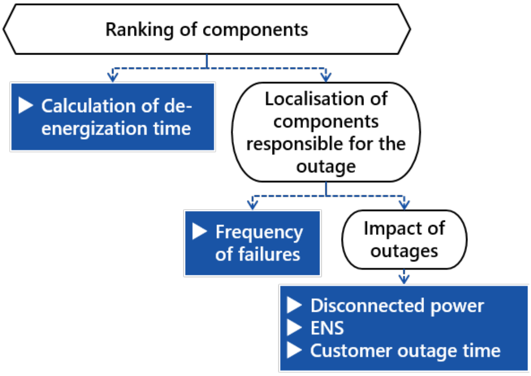


Figure 5.1: Criteria used to calculate component importance indices.

Results of component ranking for an actual substation

The results of the rankings by each of the five methods are displayed in Figure 5.2. The colors red, orange, green and blue in the right panel represent the decreasing de-energization time. The following observations can be made:

- The ranking based on the *frequency of failure* fails to provide a demarcation between the components responsible, since the range of rankings is small (from 1 to 5). This is because the frequency of failures for many components is the same. The frequency of failures does not discriminate enough between components.
- The ranking based on *de-energization time* performs better than the frequency of failure. It provides information regarding components or sections of feeders vulnerable to failures with appropriate ranking (see Figure 5.1 right panel). However, de-energization does not necessarily mean failure. When analyzing

the switch events, a set of components with high de-energization time is obtained. But only one or a few components in the set are actually responsible for the outages. In addition, components used for backup also have a high de-energization time. Hence, careful consideration is essential when interpreting the results.

- The results obtained by the ranking based on the *impact of outages* are easier to understand as they involve fewer components than the de-energization time method, while providing a sufficiently large range of rankings that enables to discriminate among components (see Figure 5.1 left panel). In addition, this ranking method takes into account the actual component at the origin of the outage.

Since each method is based on a specific criterion that is not taken into account in other methods (for example failure frequency, type and severity of impacts), the selection of the "best" method depends on the primary goals and acceptability levels of the grid operator. Besides, the components ranked high in all the methods can be identified. These components can be seen as critical, and would need a focused monitoring to prevent outages that can have high impacts for distribution system operators.

Contrary to the methods reviewed (see **Paper IV**), which are computationally expensive, require data that are not commonly available at the distribution level (for example the expected outage rate and duration of components), or use generic values for all components, the proposed ranking methods provide a simple and easily understandable way to identify critical components, using accessible empirical data. This enables to focus maintenance strategies, identify data collection needs, and develop redundancy infrastructure at identified critical points, eventually improving the reliability of the grid. This is essential at the distribution level, where a continuous monitoring apparatus covering all components, or a completely redundant infrastructure are not economically feasible.

5.2 Prediction of failures using maintenance, failure, and weather-related data

Component-level failure detection

This section is based on **Paper V**.

Automatization algorithm. The literature related to the use of data analytics, and especially machine learning, for health prediction purposes is extensive. However, the datasets used are often different from real-world datasets.

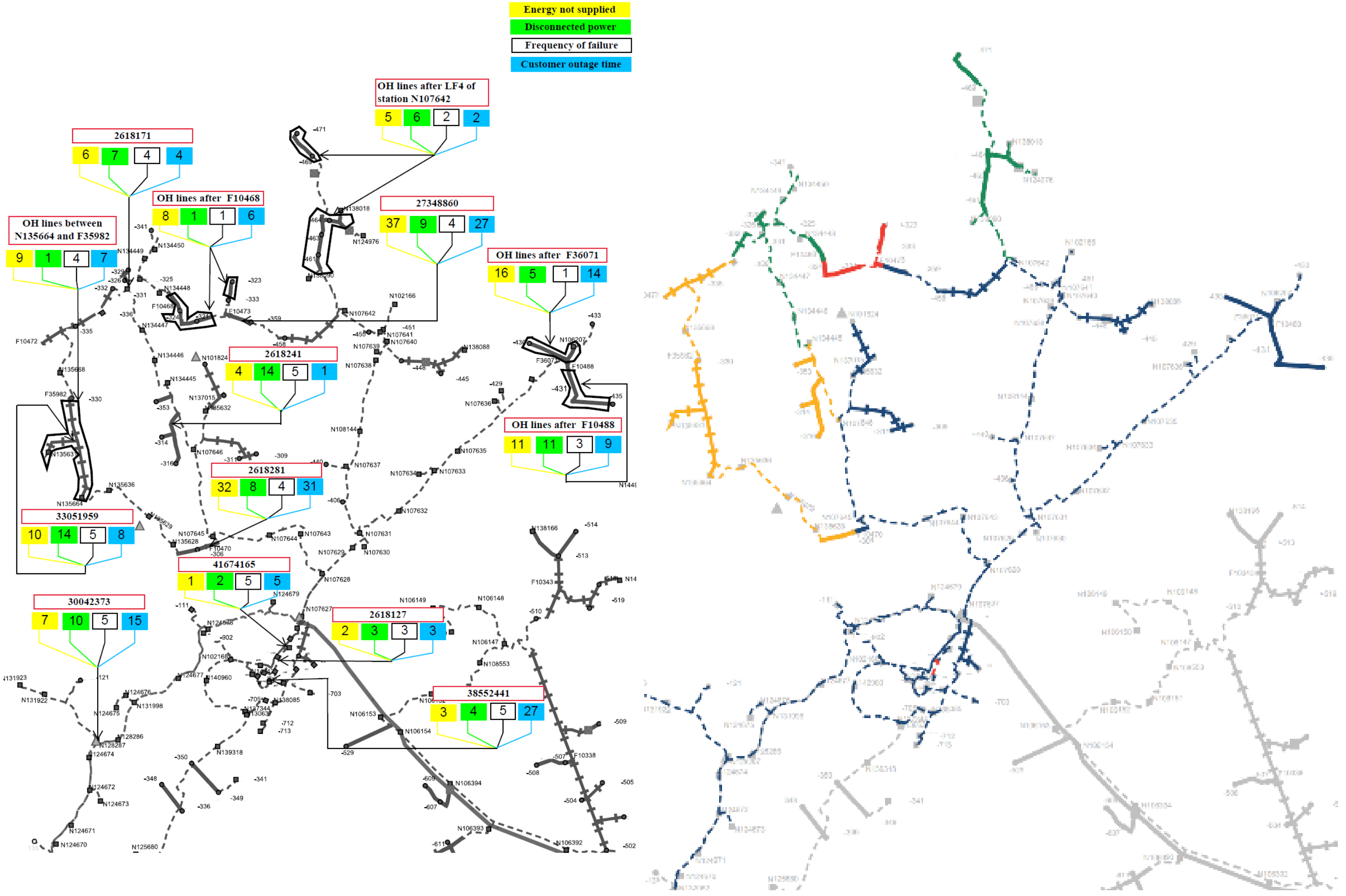


Figure 5.2: Components and their rank using the localisation (left panel) and de-energization-time (right panel) methods.

In practice, data are often multivariate, sequential[1], imbalanced[2], and have a low sampling rate. Moreover, many failure prediction algorithms use data from component-specific sensors. The more sensors with high sampling rate, the higher the quality of the dataset, and the easier the classification task. While this configuration is possible for important components, it is unrealistic in other cases, when a system is composed of many components whose failure is not life threatening, as in the case of power systems. Thus, many of the methods developed for sensor data with high sampling rate might not be useful for the prediction of component failure in practice. Another challenge is the high amount of models and hyperparameters that need to be fine-tuned. It becomes intractable to test all models and hyperparameter sets to develop individual models.

Our goal is to investigate the possibility to predict rare events with multivariate, sequential, imbalanced datasets of low sampling rate, and without component-specific sensors. We explore in which ways the usual classification methods are relevant for a dataset with the afore-mentioned characteristics.

To this purpose, we designed an automatization algorithm to automatically select optimal hyperparameters for different models. It is composed of four phases:

- The input pre-processing phase results in several datasets, used to analyze the influence of factors related to data pre-processing on model performance. The factors are as follows: i) binary variables, ii) lag numbers and sliding window, and iii) model type, and use of ensemble models;
- The hyperparameter optimization phase results in the selection of several sets of hyperparameters for each selected model and each dataset. The resulting sets of models are ranked according to the performance of hyperparameters sets on validation data;
- The evaluation phase consists in testing the selected models on the selected datasets, with a held-out test set;
- In the last phase, the performance of each model on each dataset is analyzed and comparisons are performed to identify how some characteristics of the datasets influence the results.

Six supervised learning and two unsupervised learning classification methods are selected. Four of them are classical machine learning methods, four of them are deep learning methods (neural networks):

- logistic regression (LOGIT)

[1] The order of datapoints must be taken into account because there are dependencies between them. For example time series or DNA sequences.

[2] Dataset with skewed class proportions, containing majority classes that make up a large proportion of the dataset, and minority classes that make up a smaller proportion.

- support vector machines (SVM)
- random forests (RF)
- extreme gradient boosting (XGBoost)
- multi-layer perceptrons (MLP)
- convolutional neural networks (CONV)
- autoencoders with fully connected layers (AE-MLP)
- autoencoders with convolutional layers (AE-CONV).

One particularity of autoencoders is that the minority class[2] is not needed for model training, which is advantageous when the minority event is a rare event.

Implementation on a real-world dataset. We apply the algorithm to the prediction of the failure of a high-voltage, direct current (HVDC) power line, using commonly available data. Eight variables are used in the input: i) maintenance; ii) events in nearby AC links; iii) power exchanges; iv) solar radiations; v) humidity; vi) maximum wind speed; vii) visibility; and viii) historical failures.

This is the first time a component failure prediction is attempted using only maintenance information, weather-related data, and data on neighboring AC links events. The performance results in terms of area under the curve (AUC) are displayed in Figure 5.3. Figure 5.4 shows the confusion matrix of the two best performing models.

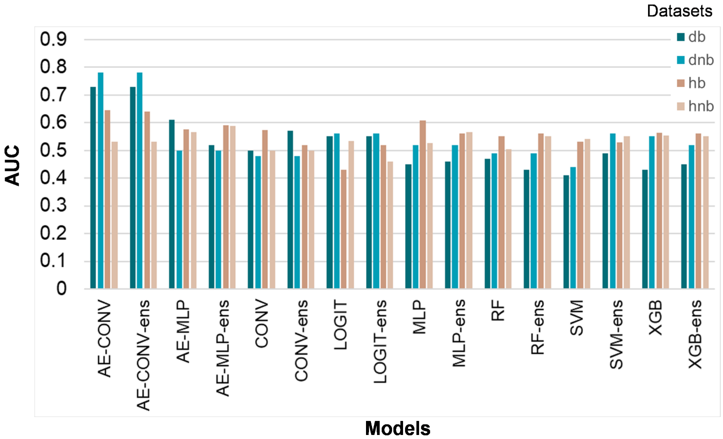


Figure 5.3: Ranking of models and associated datasets, in terms of AUC.

The main findings from **Paper V** are as follows:

- Autoencoders perform better than other models, especially, autoencoders with convolutional layers;
- While taking ensemble models did not generally improve the performance, the impact of pre-processing is significant, not only on the computing time, but also on the AUC.

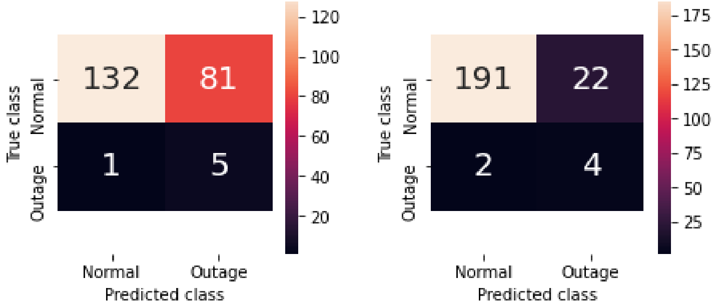


Figure 5.4: Confusion matrix of the two highest ranked models: Model=AE-CONV, dataset=*db*, AUC=0.73, F-score=0.11 (left panel) and Model=AE-CONV, dataset=*dnb*, AUC=0.78, F-score=0.25 (right panel). See **Paper V** for characteristics of the datasets.

- Substantial care should be brought to the choice of the criteria to select the best models. The AUC has the advantage of synthesizing many concepts, but another criterion like the number of false negative could be more suitable for cases where true positives (outages) must be detected and avoided at any costs.
- One difficulty in rare event prediction is the trade-off between the necessity to detect as many outages as possible (maximize the true positives), and to minimize the number of false alarms (false positives).
- Another difficulty is to predict outages that are caused by a factor that is not reflected in the input data. In the field of component failure, information about the cause of outages is scarce or not reported, which increases the difficulty to train a model.

Further improvement of the algorithm would be to synthetically generate more outage data through Generative Adversarial Networks (GAN) instead of using Synthetic Minority Oversampling Technique (SMOTE). An alternative would be to use transfer learning, which would offset the issue of outage information scarcity, by leveraging the outage experience of other similar components.

Substation-level failure detection

This section is based on [1]. See also contributions in section List of Papers.

Naive Bayes classification. At the substation level, outages happen relatively frequently, and affect several components. One approach to prevent outages from happening would be to model the failure distribution for each of the components. However, this would require data that are inexistent in practice, and would be uneconomic since the cost of installing measuring devices at all components would by far exceed the costs associated with the outages at the distribution level. Instead, the idea is to divide the substation into several areas, and study the relations between various factors or *conditions*, and the *outages* happening in each of those areas. The goal is, for a particular set of future conditions, to be able to point to an area, and a component type with the highest probability of being affected by an outage, as shown in Figure 5.5.

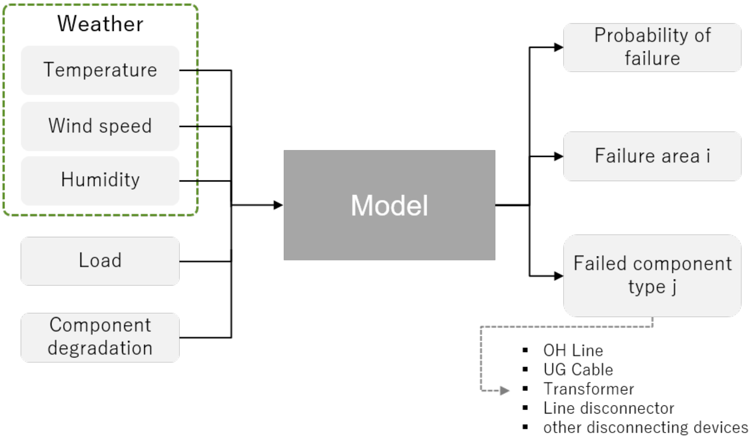


Figure 5.5: Model for predicting outages at the substation level.

The *conditions* refer to available data such as weather-related data, failure and maintenance history in the substation, load measurements and information on component degradation state (usually age of the component).

Given the small amount of data, categorical Naive Bayes has been selected as a classification method. In fact, this method tends to work well even with limited amounts of data, contrary to neural networks. Also, updating the model when new data are available is easy, because no heavy computing is involved. The aim of Naive Bayes is to obtain the probability of a hypothesis given some conditions (posterior probability $P(Hyp_h|Evi_e)$), and can be formulated as follows:

$$P(Hyp_h|Evi_e) = \frac{P(Hyp_h \cap Evi_e)}{P(Evi_e)} = \frac{P(Evi_e|Hyp_h)P(Hyp_h)}{P(Evi_e)}$$

where: $Evi_e = \{Temp_e, Wind_e, Hum_e, Load_e, Deg_e\}$ and $Hyp_h = \{Area_h, Comp_h\}$
Evi_e are the evidences or conditions
$Temp_e$ is the temperature category
$Wind_e$ is the wind speed category
Hum_e is the humidity category
$Load_e$ is the load category
Deg_e is the degradation category

Hyp_h is the hypothesis (about area and component type affected by an outage)
$Area_h$ is the failure area
$Comp_h$ is the component type.

An important step in the methodology is to categorize the continuous variables. The categorization is based on quantiles. Besides, the maximum value of windspeed and of relative humidity in the past 6 hours, as well as the average value of temperature over the 6 hours preceding the outage are calculated before being categorized using quantiles. As for the load, a percentage load is calculated considering the average value of load during the outage with respect to the monthly average. This categorization enables to create a limited number of condition sets. Any situation can thus be classified into one of the 384 possible combinations created. Among them, 80 combinations are associated with actual failures.

Implementation on a real-world dataset. For a particular condition set, the categorical Bayes classifier provides answers to the following questions: What is the probability that an outage happens? What type of component will be affected? What type of area will be affected? Moreover, it can also give information on the conditions that are critical. i.e. with higher probability of failure.

In the case of the substation under study, the probability of the outage is highest during summer (0.385), and autumn (0.266). The most affected components are OH lines and UG cables. The areas most vulnerable to failures are area 2 and area 4, which are most affected in summer and autumn, notably because of tree-falls and thunderstorms.

The results of the prediction model on a test dataset are represented in Table 5.1. The table gives seven samples of outages, the conditions associated to the outage, and the area and component actually affected. Then, it provides a list of predictions given by the model, which include the areas that could be affected by an outage, and inside each area, the type of component affected. The last column provides the probability that the outage affects the area and component type predicted. Over the seven cases, the list provided by the model included the correct area and component type that would be affected in four cases. However, the highest probabilities were assigned to other elements of the lists. Therefore, the prediction of location and component type likely to be affected is relatively inaccurate, since the number of outages contained in the dataset is low. By collecting more data, and updating the model accordingly, the predictions would be more accurate. Besides, further investigations are needed to evaluate the false positives on non-failure data.

In conclusion, the prediction model can be used to alert the distribution system operator about possible outages in the network for a given set of weather conditions. One condition for the model to be accurate is to have a long historical dataset.

Table 5.1: Prediction of the area and type of component to be affected by failures. Predictions in bold correspond to the actual area and component type affected.

Outage ID	**Input**	**Actual area & component**	**Predictions**	**Probability**
2292669	T2,W2,H3	Area 7 UG Cable	Area 2, OH Line	4/6111
			Area 3, OH Line	2/6111
			Area 4, OH Line	1/6111
			Area 7, UG Cable	1/6111
2294308	T2,W2,H3	Area 4 OH Line	Area 2, OH Line	4/6111
			Area 3, OH Line	2/6111
			Area 4, OH Line	1/6111
			Area 7, UG Cable	1/6111
2296566	T2,W2,H3	Area 3 OH Line	Area 2, OH Line	4/6111
			Area 3, OH Line	2/6111
			Area 4, OH Line	1/6111
			Area 7, UG Cable	1/6111
2298416	T2,W2,H3	Area 4 UG Cable	Area 2, OH Line	4/6111
			Area 3, OH Line	2/6111
			Area 4, OH Line	1/6111
			Area 7, UG Cable	1/6111
2297573	T3,W2,H3	Area 4 OH Line	Area 2, OH Line	3/4642
			Area 3, OH Line	1/4642
			Area 3, UG Cable	1/4642
			Area 4, OH Line	2/4642
			Area 4, Transformer	1/4642
			Area 4, UG Cable	2/4642
			Area 5, UG Cable	1/4642
2299539	T1,W1,H3	Area 2 OH Line	Area 1, OH Line	1/4192
			Area 2, OH Line	1/4192
			Area 2, Transformer	1/4192
			Area 4, OH Line	2/4192
			Area 4, UG Cable	2/4192
			Area 6, UG Cable	1/4192
2299564	T1,W1,H2	Area 1 OH Line	No outages recorded historically	0

This chapter shows that already available data can be used more efficiently, and give relevant insights to the grid operator, by using innovative approaches. Approaches within the field of selection and prediction have been developed. They can be used to prioritize components, and warn about possible outages, thus contributing to a risk-based scheduling of component replacement and maintenance.

Chapter 6

Conclusion

Main conclusions and discussion

This thesis is focused on the relations between *data* and asset management *decisions* in the power sector. The two concepts are linked together through *approaches.* An approach consists in processing the data in such a way that the result is used as a basis to take an action related to the assets. These relations are important because the efficiency of the decisions taken can directly be traced back to the kind of approach chosen and the data that have been used. The influence of data quality on the decisions has been investigated in **Paper I**, with a focus on component replacement as asset management decision. Low quality data shift the choice of year of replacement, leading to higher annual costs for the use of the component. A framework has been developed to support asset managers to decide the optimal level of data quality, which is the level that is economically profitable.

The need to reflect upon the amount and quality of data that is necessary to take efficient actions is even more pronounced since the development of machine learning and the energy transition. The multiplication of smaller-scale power generators and new loads to fight climate change is complexifying the grid, and makes the decision making process more challenging. Therefore, the transition to smart grids is closely linked to the deployment of meters and sensors, and of machine learning algorithms, which in turn supports automatization, to automatize modeling and reduce manual work. These new trends are analyzed and detailed in **Papers II and III**.

However, installing sensors on all components of the distribution system would be prohibitively expensive. Therefore, many of the recently developed algorithms, based on sensor data with high sampling rate, often cannot be used in practice. Moreover, more data collection means also increased processing needs and storage issues. It follows that data needs should be carefully evaluated, and that new approaches should first take into account the real-world state of datasets, before considering the use of data that could potentially be collected in the future.

In **Paper IV**, empirical approaches to assign an important index to components of a substation are proposed. They use only commonly available data such as the switch log, outage history, and outage impacts in terms of energy not supplied and duration. The proposed rankings would enable to identify critical components, and to adapt data and asset management accordingly.

In **Paper V**, an automatization algorithm has been developed for failure prediction, also using commonly available data. It provides a basis of reflection about the selection of algorithms that might be fit for multivariate, low frequency, sequential, imbalanced, and unspecific datasets, which are very common in the industry. It also shows that automatization can be used to reduce the need to install sensors, by using more efficiently the data that are already available, instead of supporting a race for big data.

A specific challenge common to the works presented in this thesis is the need to make choices, particularly concerning trade-offs. For example **Paper I** deals with finding a balance between the costs of increasing data quality, and the benefits of taking more efficient decisions based on data with improved quality. In **Paper V**, there is a trade-off between the precision of the prediction, and benefits of a more precise prediction. Having excellent data for a particular component is profitable when the costs of failure of this component are high. Otherwise, other solutions need to be developed, as shown in **Papers IV and V**. One way is to collect data only for important components. Another way is to ”zoom out”, which means collecting high quality data for an area, not a single component, so as to reduce the uncertainty in the area to an acceptable level. Therefore, an important step in data management is to define the precision that is acceptable for each area or situation, and then to pro-actively monitor data needs and upgrade them if necessary.

Future work

Several avenues for further research are considered, that are all based on investigating the relations between *data* and grid *performance*, as conceptualized in Figure 2.1. One avenue is to study the benefits of installing *new high sampling rate devices*, and explore whether more efficient *decisions* are taken. Another avenue is to produce a state of the art of *common measurement devices*, and examine in which ways *new approaches* using the available data would enable to achieve better decisions. Finally, an interesting avenue would be to investigate whether *current approaches and data collection level* are adapted to a change in the *situation* of the power system.

References

[1] S. K. Nalini Ramakrishna, "Component importance indices and failure prevention using outage data in distribution systems," Master's thesis, KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, 2020.

[2] M. Antonelli, U. Desideri, and A. Franco, "Effects of large scale penetration of renewables: The italian case in the years 2008–2015," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 3090–3100, 2018.

[3] Zhe Wang, Jingru Li, Weihong Yang, and Zinan Shi, "Impact of distributed generation on the power supply reliability," in *IEEE PES Innovative Smart Grid Technologies*, pp. 1–5, 2012.

[4] S. Riaz, H. Marzooghi, G. Verbič, A. C. Chapman, and D. J. Hill, "Generic demand model considering the impact of prosumers for future grid scenario analysis," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 819–829, 2019.

[5] K. Balamurugan, D. Srinivasan, and T. Reindl, "Impact of distributed generation on power distribution systems," *Energy Procedia*, vol. 25, pp. 93–100, 2012. PV Asia Pacific Conference 2011.

[6] S. Zhou and M. A. Brown, "Smart meter deployment in europe: A comparative case study on the impacts of national policy schemes," *Journal of Cleaner Production*, vol. 144, pp. 22–32, 2017.

[7] J. Zheng, D. W. Gao, and L. Lin, "Smart meters in smart grid: An overview," in *IEEE Green Technologies Conference (GreenTech)*, pp. 57–64, 2013.

[8] M. Moness and A. M. Moustafa, "A survey of cyber-physical advances and challenges of wind energy conversion systems: Prospects for internet of energy," *IEEE Internet of Things Journal*, vol. 3, no. 2, pp. 134–145, 2016.

[9] K. Grave, M. Paulus, and D. Lindenberger, "A method for estimating security of electricity supply from intermittent sources: Scenarios for Germany until 2030. The paper is based on a study of the Institute of Energy Economics at the

University of Cologne, funded by the German Federal Ministry of Economics and Technology (BMWI) which assessed German electricity supply security in the short- and mid-term.," *Energy Policy*, vol. 46, pp. 193 – 202, 2012.

[10] S. Duvnjak Zarkovic, *Security of Electricity Supply in Power Distribution System. Optimization Algorithms for Reliability Centered Distribution System Planning (Licentiate dissertation).* KTH Royal Institute of Technology, Stockholm, 2020.

[11] E. Duarte, D. Falla, J. Gavin, M. Lawrence, T. McGrail, D. Miller, P. Prout, and B. Rogan, "A practical approach to condition and risk based power transformer asset replacement," in *IEEE International Symposium on Electrical Insulation*, pp. 1–4, 2010.

[12] R. Rocchetta, L. Bellani, M. Compare, E. Zio, and E. Patelli, "A reinforcement learning framework for optimal operation and maintenance of power grids," *Applied Energy*, vol. 241, pp. 291 – 301, 2019.

[13] X. Zhao, K. N. Al-Khalifa, A. Magid Hamouda, and T. Nakagawa, "Age replacement models: A summary with new perspectives and methods," *Reliability Engineering and System Safety*, vol. 161, pp. 95 – 105, 2017.

[14] J. H. Jürgensen, L. Nordström, and P. Hilber, "Estimation of individual failure rates for power system components based on risk functions," *IEEE Transactions on Power Delivery*, vol. 34, no. 4, pp. 1599–1607, 2019.

[15] W. Li, J. Zhou, J. Lu, and W. Yan, "A probabilistic analysis approach to making decision on retirement of aged equipment in transmission systems," *IEEE Transactions on Power Delivery*, vol. 22, no. 3, pp. 1891–1896, 2007.

[16] P. Hilber and L. Bertling, "Component reliability importance indices for electrical networks," in *International Power Engineering Conference (IPEC)*, pp. 257–263, 2007.

[17] S. K. E. Awadallah, J. V. Milanović, and P. N. Jarman, "Reliability based framework for cost-effective replacement of power transmission equipment," *IEEE Transactions on Power Systems*, vol. 29, no. 5, pp. 2549–2557, 2014.

[18] G. K. Tayi and D. P. Ballou, "Examining data quality," *Commun. ACM*, vol. 41, p. 54–57, Feb. 1998.

[19] "Energy roadmap 2050 (com/2011/0885) - communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions," *OJ*, 15.12.2011.

[20] H. Xiao, Y. Huimei, W. Chen, and L. Hongjun, "A survey of influence of electrics vehicle charging on power grid," in *9th IEEE Conference on Industrial Electronics and Applications*, pp. 121–126, 2014.

[21] K. Janda, J. Málek, and L. Rečka, "Influence of renewable energy sources on transmission networks in central europe," *Energy Policy*, vol. 108, pp. 524 – 537, 2017.

[22] H. V. Poor, *An introduction to signal detection and estimation.* Springer Science & Business Media, 2013.