Doctoral theses at NTNU, 2022:308

Michael Pacevicius

Optimization of Information Management for Dynamic Risk Analysis of Large-scale Power Grids

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Engineering
Department of Mechanical and Industrial
Engineering

NTNU
Norwegian University of
Science and Technology

Michael Pacevicius

Optimization of Information Management for Dynamic Risk Analysis of Large-scale Power Grids

Thesis for the Degree of Philosophiae Doctor

Trondheim, October 2022

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

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Engineering
Department of Mechanical and Industrial Engineering

ISBN 978-82-326-5650-9 (printed ver.)
ISBN 978-82-326-6447-4 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2022:308

Printed by NTNU Grafisk senter

Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) for partial fulfilment of the requirements for the degree of Philosophiae Doctor.

The work was carried out at both the Analytical Department of eSmart Systems, in Halden, Norway, and the Department of Mechanical and Industrial Engineering at NTNU, in Trondheim, Norway. Professor Nicola Paltrinieri from the Department of Mechanical and Industrial Engineering at NTNU was the main supervisor. Dr. Davide Roverso, Chief Analytics Officer at eSmart Systems, as well as Pierluigi Salvo Rossi, Professor at the Department of Electronic Systems at NTNU were the first and the second co-supervisors respectively.

This Industrial Ph.D. project was funded by eSmart Systems and the Norwegian Research Council (NæringsPh.D. program - Project No.: 276404). The author was also involved in research activities supported by the European Space Agency (ESA) through the GridEyeS project (Contract No.: 4000127831/19/NL/MM/ra).

The target audience of this work includes researchers and practitioners interested in the following areas: risk management, power grid management, data management, systems engineering, software development. The work may also be beneficial to regulatory authorities and standardization institutions.

Michael Pacevicius

Strasbourg, 2022

Summary

Modern societies strongly rely on the reliable functioning of power grids. However, the management of such infrastructures is a challenging task. It requires the existence of efficient solutions enabling adequate decision making, both during normal working conditions and in emergency situations. Unfortunately, blackouts and other large-scale outages continue to be regularly observed over the globe, impacting millions of people with sometimes fatal consequences. One reason for the occurrence of such events is the inadequate capture and processing of data, which has hindered the development of performant risk-focused tools to be used in practice by utilities. This can be observed in the management of vegetation along power lines, a common source of disturbances that has contributed to the occurrence of multiple well-known high-impact outages.

The digitalization of our society represents an opportunity for more accurate, data-informed risk analysis, as it supports facilitated access to more and better data. However, there is still a lack of standards, guidelines, and recommendations indicating how the data can be processed in the field of power grid management to reduce the probabilities and consequences of undesired, critical events.

The present Ph.D. addresses this gap by investigating both the fields of risk analysis and power grid management. It starts by taking advantage of the strong industrial environment in which this Ph.D. has been executed to adequately identify the relevant stakeholders and understand their needs and constraints. This is particularly important, as it directly impacts the development and thus the performance of tools to be used in the future by power grid managers. The thesis also determines, after an intensive literature review, which gaps need to be addressed in the field of risk analysis to enable efficient, large-scale heterogeneous data processing.

The main contributions of this thesis can be summarized as follows:

(1) We diversify the panel of exploitable data sources for risk analysis and fully explore the analysis level scale.
(2) We augment conventional risk assessment frameworks to enable efficient, large-scale heterogeneous data processing.
(3) We provide multiple solution development propositions enabling power grid operators to make better risk-based decisions. The propositions are based on various perspectives and enable finding an adequate trade-off between global and local analyses of the grid, by always keeping the user-needs at the center of the solution definition.
(4) We make multiple recommendations usable by power grid operators to optimize the exploitation of historical data and the planning of future data capture.
(5) We use the four previously reported contributions to indicate how vegetation management along power lines may be improved

From a risk perspective, this Ph.D. first contributes to the understanding and clarification of basic risk-related concepts. The findings of this work then enable risk analysis processes to better leverage accessible data sources. Those especially enable more robust decision-making by reducing uncertainties relative to data integration, therefore “better knowing how well we

know”. In addition, the results strongly contribute to a better quantification of problems at scope in risk analysis, de facto enabling more objective decision-making.

The thesis is also particularly valuable from a power grid management perspective. It first provides a familiarization opportunity with the notion of risk for the stakeholders requiring further insights in that field. It then shows how this knowledge can be used in combination with news data capture and processing solutions to enable the emergence of innovative tools supporting power grid operators in their daily operations.

The final results are discussed, and different evolution opportunities are reported along with the provided contributions, such as executing risk quantification or analyzing other hazards. The provided suggestions represent as many possibilities to reinforce and further extend the results of this doctoral project. They are also an indication that further development is required to facilitate more robust decision-making when practical implications and currently existing technical limitations are faced.

This thesis is a good illustration of the benefits of braking silos and encouraging cross-disciplinary cooperation. It stimulates power grid operators to further investigate the advances made in the academic world. At the same time, it also favors the communication of constraints faced in real-world situations but maybe too often excluded from the research scope in fundamental research.

Acknowledgment

This Ph.D. dissertation closes over four years of an intensive journey. As most Ph.D. candidates at this stage of the doctoral project, I am happy to conclude the redaction of this thesis by taking the time to thank those that enabled me to make this dream come true.

I would first like to sincerely thank my three supervisors, Professor Nicola Paltrinieri, Dr. Davide Roverso and Professor Pierluigi Salvo Rossi for believing in this project. By accepting and supporting the initiation of this project as I first presented it in 2017, you gave me the opportunity to do exactly what I wanted to do: Go to Norway and start an industrial Ph.D. project mixing IT, data mining and risk analysis on a topic strongly relevant for modern societies - power grid management. This was far from being a given!

Nicola, thank you for your continuous support over the years! Your experience, your knowledge and your achievements are both impressive and a source of inspiration for any young researcher. Your way of challenging my work was a great opportunity for growth. You have taught me a lot of what I know in the field of risk, but also about research and scientific rigour. The freedom that you then gave me has allowed me to explore areas which I believed were interesting and meaningful. Thank you as well for suggesting a good guide for the PSAM conference in Los Angeles! I will always remain in debt.

Davide, the sharpness of your thoughts and your incredible knowledge in the field of data management have been both an opportunity to challenge myself and a continuous source of motivation for learning more. Thank you as well for your trust and your support during the tough times of this doctoral project, which have been real sources of encouragement and which have enabled me to start this new life.

Pierluigi, your support and your advices have allowed me to enlarge the horizon of possibilities in a research area critical for the thesis. Thank you so much for that as well.

I would also like to acknowledge Knut H. H. Johansen, former CEO of eSmart Systems as well as the members of the Norwegian Research Council, who have enabled and strongly supported this project.

I would like to thank my co-authors for our constructive exchanges and for their contributions in the articles we prepared. I also take this opportunity to thank my colleagues at both eSmart Systems and at the Department of Mechanical and Industrial Engineering at NTNU, as well as the numerous stakeholders I had the chance to work with during the different projects I have been involved in. They have supported me with my research activities by enabling data access, processing and understanding, but have also been of great support from a personal perspective. Our coffee breaks, lunches, sport sessions and discussions have been particularly pleasant in the sometimes-challenging periods of this Ph.D. journey.

A special thanks to my officemate Federico, whose friendship, amazing creativity and life visions have been incredibly valuable in the most demanding periods of the thesis.

I would also like to sincerely thank my parents, siblings and other family members; as well as my other friends spread over the world for their love, their curiosity and their encouragements. Their emotional support has allowed me to face the toughest moments of this Ph.D. with strength, calm and philosophy.

Last – and foremost – I would like to thank my spouse, Marilia. Your love, your joy and your support are the reasons I could face the incredible sacrifices we had to make in this crazy period. I am happy to know that the last words of this thesis are those that have enabled me to get back to you.

Michael Pacevicius

Strasbourg, 2022

Table of Content

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANNs	Artificial Neural Networks
API	Application Programming Interface
CD	Connected Drone
CE	Critical Event
DNNs	Deep Neural Networks
DSOs	Distribution System Operators
ESA	European Space Agency
ET	Event Tree
FT	Fault Tree
IDC	International Data Corporation
KPIs	Key Performance Indicators
MLR	Multiple Linear Regression
QRA	Quantitative Risk Analysis
R&D	Research and Development
RoI	Regions of Interest
ROW	Rights of Way
SE	Systems Engineering
TSOs	Transmission System Operators

Thesis Structure

This Ph.D. thesis has been prepared as a collection of articles, which is nowadays the most common way of preparing a thesis at NTNU. The dissertation consists in a compilation of the different research activities that have been executed and reported for dissemination during the doctoral project.

This thesis consists of two main parts: *Part I – Main Report* and *Part II – Articles*.

Part I – Main Report

The first part summarizes the research activities executed during the thesis. It highlights the main objectives, results, and contributions and shows how the mentioned articles are interrelated.

Part II – Articles

The second part consists of 10 articles submitted to international peer-reviewed conferences or journals. Nine have been published or accepted, and one is currently under review. Those articles represent the backbone of the present thesis, are stand-alone, and can be read in any order.

First, considering Part I enables obtaining a complete overview of the doctoral project, while Part II goes deeper into the research details. Therefore, I also suggest this order for the reading of the thesis. However, both parts are stand-alone, and it is up to the reader to decide about the most convenient reading order based on their preferences.

Declaration of Contributions

The present Ph.D. dissertation is based on the journal articles and conference articles listed in Table I. The author's contribution is also detailed in this section, while the full version of the reported articles is available in Part II of the thesis.

Table I – Overview of articles included in the thesis.

Article N°	Conference (C) or Journal (J)	Title
I	ESREL 2018 (C)	Risk of crack formation in power grid wooden poles and relationship with meteorological conditions: a Norwegian case study
II	PSAM 2018 (C)	Smart Grids: Challenges of Processing Heterogeneous Data for Risk Assessment
III	CSER 2020 (C)	Supporting the Application of Dynamic Risk Analysis to Real-World Situations using Systems Engineering: A focus on the Norwegian Power Grid Management
IV	ESREL-PSAM 2020 (C)	Optimizing Technology-based Decision-support for Management of Infrastructures under Risk: The Case of Power Grids
V	ESREL-PSAM 2020 (C)	Lessons from past hazardous events: data analytics for severity prediction
VI	IEEE Transactions on Power Delivery 2021 (J)	Automated Power Lines Vegetation Monitoring using High-Resolution Satellite Imagery
VII	Energies 2022 (J)	Managing Heterogeneous Datasets for Dynamic Risk Analysis of Large-Scale Infrastructures
VIII	ASME J. Risk Uncertainty Part B. 2022 (J)	Data-informed Risk Analysis of Power Grids: Application of Method for Managing Heterogeneous Datasets
IX	RAMS 2021 (C)	Heterogeneous Data-merging Platform for Improved Risk Management in Power Grids
X	RAMS 2021 (C)	Addressing the Importance of Data Veracity during Data Acquisition for Risk Assessment Processes

Article I:

Pacevicius, M., Roverso, D., Salvo Rossi, P., Paltrinieri, N., 2018b. **Risk of crack formation in power grid wooden poles and relationship with meteorological conditions: a Norwegian case study**, in: European Safety and Reliability Conference - ESREL 2018. Trondheim, p. 7.

I initiated the research idea, defined the research approach, and conducted the literature review. I wrote the manuscript draft, handled the submission process and presented the work. The co-authors assisted in structuring the work and refined the manuscript.

Article II:

Pacevicius, M., Roverso, D., Salvo Rossi, P., Paltrinieri, N., 2018a. **Smart Grids: Challenges of Processing Heterogeneous Data for Risk Assessment**, in: Probabilistic Safety Assesment and Management - PSAM 14. Los Angeles, p. 11.

I initiated the research idea, defined the research approach and conducted the literature review. I also handled the investigations presented in the work. More significantly, I assessed which data sources could be considered and how they could be combined. Finally, I wrote the manuscript draft, handled the submission process and presented the work. The co-authors assisted in structuring the work and refined the manuscript.

Article III:

Pacevicius, M., Haskins, C., Paltrinieri, N., 2020. **Supporting the Application of Dynamic Risk Analysis to Real-World Situations using Systems Engineering: A focus on the Norwegian Power Grid Management**, in: 18th Annual Conference on Systems Engineering Research - Recent Trends and Advances in Model-Based Systems Engineering. Virtual, p. 8.

I initiated the research idea, defined the research approach and conducted the literature review. I also handled the investigations presented in the work based on the GridEyeS project, which I co-managed. I wrote the manuscript draft, handled the submission process and presented the work. The co-authors provided feedback on the approach, analysis and arguments. They also assisted in structuring the work and refined the manuscript.

Article IV:

Pacevicius, M., Ramos, M.A., Paltrinieri, N., 2020. **Optimizing Technology-based Decision-support for Management of Infrastructures under Risk: The Case of Power Grids**, in: 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference. Venice, p. 8.

I initiated the research idea, defined the research approach together with the second author and conducted the literature review. I also handled the discussions with grid operators and assessed the informative potential of the considered data sources. I wrote the manuscript draft, handled the submission process and presented the work. The co-authors provided feedback on the approach and arguments. They helped to reinforce the contribution of the work, assisted in structuring the work and refined the manuscript.

Article V:

Paltrinieri, N., Patriarca, R., Pacevicius, M., Salvo Rossi, P., 2020. **Lessons from past hazardous events: data analytics for severity prediction**, in: 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference. Venice, p. 8.

I supported the co-authors through discussions on the use of machine learning and the interpretation of the results. I assisted in structuring the work and refined the manuscript.

Article VI:

Gazzea, M., Pacevicius, M., Dammann, D.O., Sapronova, A., Lunde, T.M., Arghandeh, R., 2021. **Automated Power Lines Vegetation Monitoring using High-Resolution Satellite Imagery**. Trans. Power Deliv. p. 9.

I co-defined the research approach and dataset association possibilities. I prepared the LiDAR data used in the analytics of the work, helped to prepare the satellite imagery and advised on the usable tools. I also handled the discussions with the grid company involved in the study. This study is part of the GridEyeS project, which I co-managed. The main author developed the machine learning models and wrote the manuscript draft - which I then fully reviewed. The other co-authors provided feedback on the approach and arguments. They helped to reinforce the contribution of the work, assisted in structuring the work and refined the manuscript.

Article VII:

Pacevicius, M., Ramos, M., Roverso, D., Thun Eriksen, C., Paltrinieri, N., 2022a. **Managing Heterogeneous Datasets for Dynamic Risk Analysis of Large-Scale Infrastructures**. Energies. p. 40.

I initiated the research idea, defined the research approach and conducted the literature review. I proposed the integration of the two reinforcement actions suggested in the article and developed in addition the proposed method. I designed the case-study and performed the application of the approach. I wrote the manuscript draft and handled the submission. The last co-author advised during the development of the approach and provided feedback on the structuring of the work. The work was further evolved through inputs from the other co-authors, who also refined the manuscript.

Article VIII:

Pacevicius, M., Ramos, M., Roverso, D., Thun Eriksen, C., Paltrinieri, N., 2022b. **Data-informed Risk Analysis of Power Grids: Application of Method for Managing Heterogeneous Datasets**. ASME J. Risk Uncertainty Part B. p. 44.

I designed the case-study, structured the work and performed the application of the approach proposed in Article VII. I reviewed the available data sources, selected the usable software, and identified 17 datasets that could be used for the case-study. I processed the datasets and performed the analysis. I designed and prepared the excel file attached to the work. I wrote the manuscript draft and handled the submission. The last co-author advised during the development of the approach and provided feedback on the structuring of the work. The work was further evolved through inputs from the other co-authors, who also refined the manuscript.

Article IX:

Pacevicius, M., Dammann, D.O., Gazzea, M., Sapronova, A., 2021a. **Heterogeneous Data-merging Platform for Improved Risk Management in Power Grids**, in: The 67th Annual Reliability and Maintainability Symposium. Orlando, p. 7.

I initiated the research idea, defined the research approach together with the second author and conducted the literature review. The article is based on the GridEyeS project, which I co-managed. I wrote the manuscript draft, handled the submission process and presented the work. The strategy was mainly discussed with the second co-author, who also assisted in structuring the work and refined the manuscript. The other co-authors provided feedback on the approach, analysis and arguments.

Article X:

Pacevicius, M., Paltrinieri, N., Thieme, C.A., Salvo Rossi, P., 2021b. **Addressing the Importance of Data Veracity during Data Acquisition for Risk Assessment Processes**, in: The 67th Annual Reliability and Maintainability Symposium. Orlando, p. 7.

I initiated the research idea, defined the research approach together with the second author and conducted the literature review. I wrote most of the manuscript draft, handled the submission process and presented the work. The strategy was discussed with the second co-author, who also wrote sub-sections 2.1, 2.2 and 2.3. He also assisted in structuring the work and refined the manuscript. The other co-authors provided feedback on the approach, analysis and arguments.

Part I – Main report

1. Introduction

The ever-growing power demand of modern societies requires reliable infrastructures. Power grids represent core elements to that extent and transport energy from power producers to power consumers. However, the management of power grids is challenging for multiple reasons. For instance, power grids correspond to geographically extended infrastructures. Therefore, they are exposed to a variety of hazards, such as weather, vegetation in the proximity of the power lines, terrain instability or pests. In addition, utilities operate under constant societal and economic pressure. They must constantly identify the strategy providing the best trade-offs between different Key Performance Indicators (KPIs), such as satisfaction of the dynamic power demand, reinvestments in infrastructures and compliance with regulations – among others (Brown and Spare, 2004; Sand et al., 2007).

The importance of power grids for our societies has motived numerous research activities (See for instance Bell et al., 2012; Ciapessoni et al., 2016; Doostan et al., 2019; Guikema et al., 2006; Johansson et al., 2011; Radmer et al., 2002; Sand et al., 1989; Wanik et al., 2017, 2015). The European Union has also financed multiple projects in the last decade to reinforce the collaborations around the topic of power grid management (e.g. *AFTER*, *Umbrella*, *iTesla*, *HyRiM, Garpur)* (European Commission, 2014a, 2015, 2016, 2017a, 2017b). Supporting this type of projects enables to reinforce international cooperation for the management of an across-country connected infrastructure. Such projects also enable sharing experiences and expertise, increasing resilience when facing unexpected events.

Risk-based solutions have regularly been recommended as a conclusion of such research activities (Ciapessoni et al., 2016; Nordgård, 2010; Perkin, 2018; Vefsnmo et al., 2015). Conceptually, such solutions should optimize resource management by indicating the most critical tasks for which an action is required.

The pertinence of risk-based solutions relies on three main elements to be usable in practice (Nordgård, 2010):

(1) **Data.**
Risk-based solutions require data to understand the grid's behavior and monitor it in the long run.

(2) **Dynamic methods.**
Risk-based solutions need to integrate dynamic methods to cope with internal and external context modifications potentially requiring structural changes in the risk analysis.

(3) **User-need acknowledgment.**
Risk-based solutions need to be engineered in a way that optimally suits the power grid operator's needs.

The digitalization of our societies and the regular emergence of technologies unlock new ways of capturing data are thus critical steps toward risk-based solutions. In addition, the rise over the last decade of Dynamic Risk Analysis (DRA) methods has increased the opportunities for developing such solutions.

However, there is still a lack of adequate frameworks to make risk-based solutions a common practice in power grid management. Numerous questions remain unanswered regarding, for instance, data selection. Which decision criteria shall be considered when accessing a large panel of heterogeneous datasets? In which circumstances should such criteria be applied? Should they be identically applied to all datasets in any circumstances? How should conflicting information be managed?

In addition, the DRA methods remain in an embryonic phase, especially regarding applications in real-world configurations. The practical constraints that can be faced in such situations must be acknowledged to adequately support the development of tools and platforms that can be used for power grid operations.

The present thesis addresses those problems. It aims in particular at:

(1) Reinforcing the state-of-the-art for risk analysis applied in a real-world industrial context by exploiting the benefits of digitalization.

(2) Contributing to the development of adequate solutions enabling to support power grid operators with the decision-making processes relative to risk reduction and further resource management optimization.

To achieve these objectives, the thesis focuses principally on the management of over-head power lines. It investigates more specifically how vegetation management along such infrastructures could be facilitated in a Norwegian environment. Vegetation along power lines represents a well-known sources of outages (Doostan et al., 2019; Eggum, 2019; Hansen, 2018) and has been reported as a main contributing factor in various major blackouts in the world (Haes Alhelou et al., 2019; Sforna and Delfanti, 2006; U.S.-Canada Power System Outage Task Force, 2004). The thesis explores how heterogeneous data sources can be used to address this hazard. It does so by first benchmarking the existing technologies and data processing technics using a cross-disciplinary mindset. It then assesses and addresses the requirements for exploiting the identified data sources with conventional risk analysis. The thesis then examines how practical constraints influence the development of software solutions aimed to be implemented in control rooms; It finally provides recommendations for both the development and exploitation of the suggested solutions.

The remainder of the first part of this thesis (Part I) is organized as follows**.** **Section 2** provides the relevant research background. It recalls information essential for the understanding of *Risk Analysis* in general and reports how the topic is considered in the field of *Power Grid Management*. It also provides a high-level overview of the information commonly considered for risk analysis in power grid management. **Section 3** explicitly formulates the research challenges and adequately frames the resulting research questions around which the present thesis is built. The research objectives defined on the basis of those questions are reported in **Section 4**. **Section 5** informs about the context in which the research was executed and describes the working configuration of the thesis. **Section 6** describes the research approach that was chosen for achieving the work. **Section 7** reports the main contributions of the Ph.D. by summarizing the research findings of the articles provided in the second part of this thesis (Part II). Those contributions are discussed in **Section 8**. The conclusions and suggestions for further work are finally reported in **Section 9**.

2. Research Background

There is a large body of literature in both the fields of risk analysis and power grid management. In addition, various disciplines have been explored to generate, acquire, pre-process, analyse and compare the multiple datasets that have been used for the present thesis. The intention of the present section is not to provide a literature review of each of those fields. Instead, it aims to provide the reader with sufficient background knowledge in risk analysis and power grid management to understand which research gaps have been identified and addressed in the present Ph.D. The reader is directed to section 6 for a further detailing of additional methods that have been considered for this work and to the articles in Part II of this manuscript for more details on how those methods have been applied.

2.1.Risk management, risk assessment, and risk analysis

Risk management is a broad concept that can be addressed in multiple ways, depending on the field in which it is applied (Aven, 2012). Despite some nuances, there is general agreement that *Risk* is a function of three main elements: (1) a scenario s that could happen, (2) the probability p of this scenario happening, and (3) the consequences c of this scenario happening (Eq. 2-1) (Kaplan and Garrick, 1981).

$$Risk = f(s, p, c) \qquad \text{(Eq. 2-1)}$$

Risk management consists of a combination of three directly linked processes (Raussand, 2011):

(1) **Risk analysis.**
This consists of characterizing each element of the risk definition to obtain a relevant risk picture (i.e., listing of hazards and identification of critical events to define the relevant scenarios, definition of the related probabilities of occurrence, definition of the related consequences).

(2) **Risk Evaluation** .
This consists of generating a judgment of the generated risk picture (e.g., by comparing it with some reference criteria) and making one or several propositions to reduce the risk level.

(3) **Risk control.**
This consists in selecting and implementing the most relevant proposition provided in the risk evaluation based on pre-defined context-relevant objectives. It also consists of assessing the impact of the implemented measures and communicating the results. Finally, the outcome of this process shall cyclically be fed back to the previous processes to inform future reanalyses.

These concepts are summarized in Figure 2-1, adapted from (Raussand, 2011). Figure 2-1 also shows that the combination of the risk analysis and risk evaluation processes can be defined as "Risk Assessment".

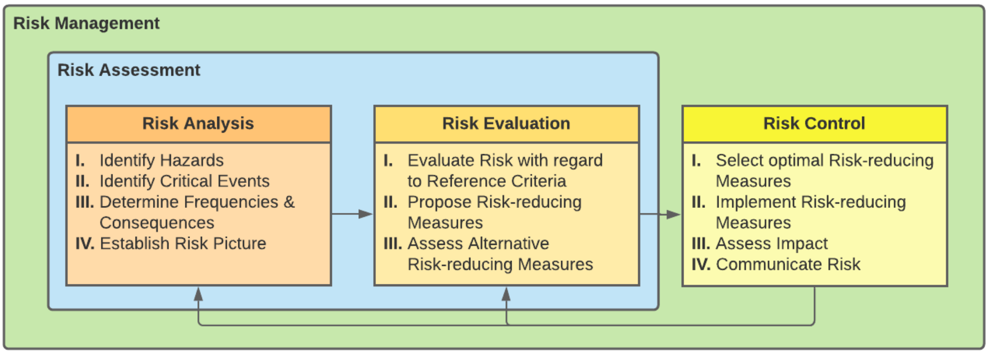
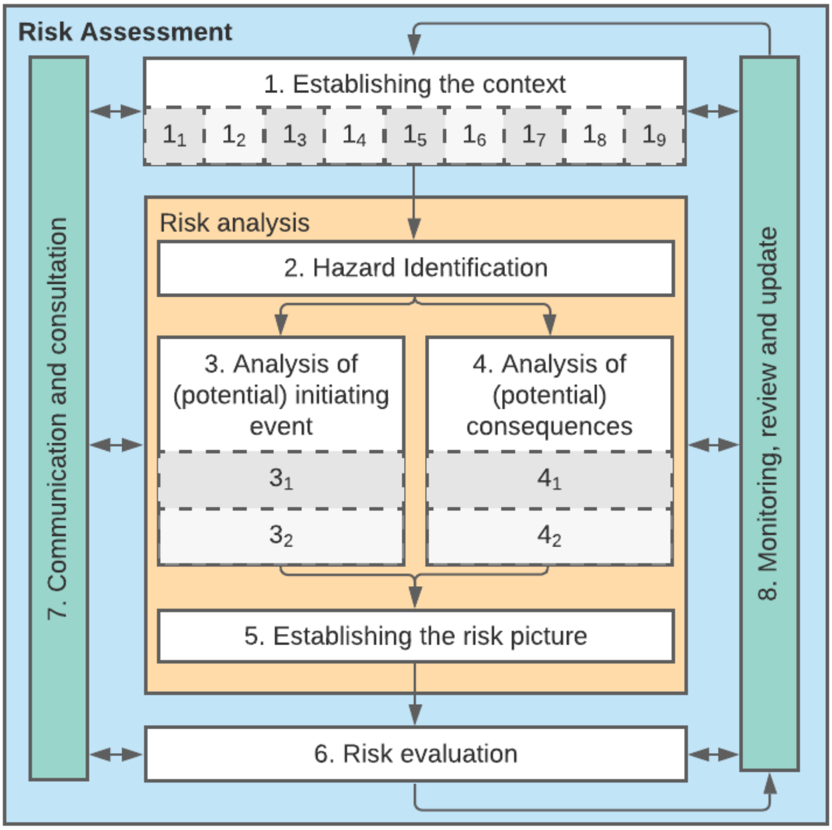


Figure 2-1: Linking between Risk Management, Risk Assessment and Risk Analysis (adapted from (Raussand, 2011))

Multiple standards relative to risk assessment are found in the literature (e.g., CSA Q850-97 (Canadian Standards Association (CSA), 1997), ISO 31000:2018 (ISO - International standardization organization, 2018); Risk governance framework (IRGC - International Risk Governance Council, 2009); NORSOK Z-013 - Risk and emergency preparedness assessment (NORSOK, 2010)). The process of the standard NORSOK Z-013 (NORSOK, 2010) established by the Norwegian petroleum industry is reported in Figure 2-2.



Detailing Step 1:

Definition of

1_1) Objectives

1_2) Scope

1_3) Responsibilities

1_4) Methods, models & tools

1_5) System boundaries & System basis

1_6) Risk acceptance criteria

1_7) Deliveries

1_8) Execution plan

1_9) Any other relevant activity

Detailing Step 3 & 4:

Identification, for initiating events, of:

3_1) Causes

4_1) Outcomes

Assessment, for initiating events, of:

3_2) Frequencies

4_2) Event sequences, Performances of satefy barriers & Magnitude/extent of the event

Figure 2-2: Description of the Risk Assessment process as defined by the standard NORSOK Z-013 (adapted from (NORSOK, 2010)). The first step consists in adequately defining 9 different points as reported at the bottom of the figure. The causes (Step 3) and consequences (Step 4) of the critical events relative to the hazards under review are also further detailed.

The standard NORSOK Z-013 is chosen as a core reference in the present thesis for the following reasons:

- The description of the risk assessment in this standard is clear and systematic.
- This standard is aligned with others, as it is based on ISO 31000 (ISO - International standardization organization, 2018).
- The oil & gas industry is a leading industry in preparing risk-related standards in Norway. As such, experience and knowledge acquired in this industry should be considered as a reference for other sectors.

In addition, the standard NORSOK Z-013 has also been used as a reference in more recent works focusing on Dynamic Risk Analysis (DRA) (Villa et al., 2016). The need for DRA solutions has emerged in the last decades and been demonstrated by the occurrence of major accidents for which the realized scenarios were initially considered as improbable, and thus ignored (Paltrinieri et al., 2012). The main conclusion in post-accident analyses was thus that the conventionally implemented risk analysis processes were not capable of acknowledging the impact of structural modifications that occurred in the operational phase (Kalantarnia et al., 2009; Pasman and Reniers, 2014).

DRA addresses this issue by focusing on the adaptability requirement of the methods. The nature of risks can drastically change over an infrastructure's lifetime due to internal and external context modifications (ISO - International standardization organization, 2018). Organizations must thus cope with such risk modifications by potentially reconsidering the tools used to manage them. Empowering this adaptability relies on both (1) the existence of data sources capable of informing the risk analysis under review and (2) being able to reconsider assumptions and models retained in previous cycles of the assessment if new risk evidence calls for updates (Kalantarnia et al., 2009; Khakzad et al., 2012; Paltrinieri et al., 2013b, 2013a; Paltrinieri and Khan, 2020).

Such methods enable, thus, theoretically, to optimize future restructuration of the risk analysis while minimizing the required reshaping efforts (Bucelli et al., 2018). However, DRA remains in an embryonic phase (Lee et al., 2019; Paltrinieri and Khan, 2020; Yang et al., 2018). Although DRA methods aim to enable adaptive responses over the lifetime of a system, there is paradoxically still a lack of feedback on the applicability of such method in the exploitation phase of the analyzed systems. The recency of both these methods and the digitalization of the society enabling a data source diversification has hindered the development of solutions usable in operation in the past and call thus for further investigations in that direction.

2.2.Power grid management and risk analysis in power grids

Power grids are sometimes referred to as the word's largest machines (Motter et al., 2013). They consist of a combination of large and small power producers of different types, numerous types of power consumers and thousands of substations – all linked with each other using power lines. The size of the grid exposes it to a plurality of hazards that can cause large blackouts. A blackout can be defined as a total or almost total absence of voltage in the transmission grid (i.e., complete collapse of the network), with consequences on a supra-regional level (European Network of Transmission System Operators for Electricity (ENTSO-E), 2010). Blackouts have in fact relatively regularly been observed all over the world, leaving thus millions of people without power (e.g., Canada (1998): 3 million; France (1999): 3,4 million; North America (2003): 50 million; Italy (2003): 45 million; Moscow (2005): 10 million; USA (2011): 6,69 million; India (2012): 620 million; USA (2012): 8,66 million; Ukraine (2015): 230 million; USA (2017): 6 million; Sudan (2018): 41,5 million; Brazil (2018): 10 million; USA (2021): 5 million (Haes Alhelou et al., 2019; Isaias Task Force, 2020; NERC, 2018; Sforna and Delfanti, 2006; Sullivan and Malik, 2021; Task Force on Power Outages - Eurelectric, 2006; U.S.-Canada Power System Outage Task Force, 2004; United States Department of Energy, 2013; Yu and Pollitt, 2009).

Power grid management is mainly the responsibility of two types of stakeholders: Transmission System Operators (TSOs) and Distribution System Operators (DSOs). TSOs usually manage grids with a voltage level above 100 kV (European Network of Transmission System Operators for Electricity (ENTSO-E), 2019). They are responsible for the power exchanges on an international level. They are also especially responsible for power transport on a national level, from the main production plants (e.g., dams, nuclear power plants or coal power plants) down to significant power-consuming endpoints (e.g., big factories) and down to the regional grid. DSOs then take over the power distribution down to smaller power consumers such as small and medium enterprises or private households. Therefore, the grid's voltage managed by DSOs is usually lower (i.e., below 100 kV), the majority of which is even below 1 kV (Eurelectric, 2020; U.S.-Canada Power System Outage Task Force, 2004).

The activities of TSOs and DSOs are particularly challenging for multiple reasons. For instance, the frequency of Extreme Weather Events (EWE) is expected to continuously increase in the future due to climate changes (Masson-Delmotte et al., 2018; Smith, 2021). This increases the risk level in an aging power grid as it increases both the probability and the consequences of damageable events (Libertson, 2021). Utilities are also continuously set under pressure regarding various socio-technical criteria. Grid operators must constantly find the right trade-off between different Key Performance Indicators (KPIs), such as satisfaction of the dynamic power demand, reinvestments in infrastructures and compliance with regulations – among others (Brown and Spare, 2004; Sand et al., 2007). These sometimes-conflicting objectives turn the network management into a complicated task, which has led to inadequate decision-making in the past. Inadequate reinvestments and maintenance strategies have been pointed out as important influencing factors in large-scale outages (FERC and NERC, 2011; Sforna and Delfanti, 2006; U.S.-Canada Power System Outage Task Force, 2004).

Numerous research activities have addressed the various aspects impacting the reliability of power grids (see for instance Bell et al., 2012; Ciapessoni et al., 2016; Doostan et al., 2019; Guikema et al., 2006; Johansson et al., 2011; Radmer et al., 2002; Sand et al., 1989; Vefsnmo et al., 2015; Wanik et al., 2017, 2015). Different projects have also obtained a financial support from the European Union over the last decade (e.g. *AFTER*, *Umbrella*, *iTesla*, *HyRiM, Garpur)* (European Commission, 2014a, 2015, 2016, 2017a, 2017b). Those projects aimed to reinforce the collaborations around power grid management and favor knowledge sharing in an international configuration.

Many of the research conclusions converge around the need to move towards a probabilistic-oriented risk assessment of power grids (Ciapessoni et al., 2016; Nordgård, 2010; Perkin, 2018; Vefsnmo et al., 2015). However, there is also an acknowledgment that this requires further research activities. Some recommendations regarding data and models were, for instance, provided at the end of the GARPUR project (GARPUR Project, 2017). Among them:

- Collect more asset-specific data to improve currently applied models.
- Integrate new types of data anytime this can be done in a cost-efficient way; evaluate how those data can be combined with existing datasets.
- Better anticipate future challenging situations; consider both large-scale and location-specific scenarios.
- Enlarge the knowledge basis with regards to new technologies.
- Enlarge both the quality and the quantity of data to gather.
- Determine the value of potentially unreliable data sources.
- Define frameworks and guidelines for collecting, maintaining, and sharing data and models.
- Explore automatization possibilities related to risks using machine learning techniques.
- Sustain a gradual development, implementation, and testing of methods relevant to a framework's definition, rather than waiting for a final holistic solution to be proposed.
- Develop new tools and sustain the formation of cross-disciplinary experts capable of providing risk-oriented analyses of heterogeneous datasets.

Nevertheless, there is still a lack of guidelines (Khuntia et al., 2017; Nordgård, 2010; Perkin et al., 2017; Vadlamudi et al., 2016) enabling to understand specifically:

- which data type can be considered as relevant,
- how this data should be assessed when different datasets are compared before being considered in a risk analysis,
- how the solutions shall technically be developed.

2.3.Information management for risk analysis in power grids

Ensuring that data can be considered as valuable for a risk analysis requires first to adequately determining who the end-users are, so that their needs can correctly be identified. In the case of power grid management, three types of operators can schematically be reported:

(1) **Grid operators in control rooms.**
Those operators focus on real-time to short-term management of the grid. Their main tasks are basically to stabilize fluctuations and ensure in real-time that the power demand can be met. This implies anticipating planned outages and trying to minimize the impact of unexpected disturbances.

(2) **Maintenance operators.**
Those operators focus on short-term to middle-term management of the grid. Their

main tasks are basically to plan and handle the maintenance operations required to maximize the uptime of the grid.

(3) **Planning operators.**
Those operators focus on middle-term to long-term management of the grid. Their main tasks are to plan and supervise the modification and extension projects of the grid. Such projects can, for instance, be motivated by the need to anticipate increases in future power demands.

The temporal rate differences of the variables involved in each of the operator tasks directly determine the type of data source that may be relevant. Grid operators in control rooms must be reactive and rely, thus, essentially on data sources providing information in near-real-time. They will mostly rely on load data and weather information. Historical and real-time records are available and used to understand the current power demand and predict coming variations. The information used by maintenance operators consists of periodic reliability assessments of the components, helping to define the infrastructure's operationality level (Catrinu and Nordgard, 2011). Planning operators finally build on the return of experience regarding the effective reliability of the components, maps enabling to characterize the environment where power grids may be installed and information regarding the nature of the future power demand (e.g., city extension, new industries) to plan the future grid modifications.

Data type selection for the operators' decision-making mainly results from practical constraints, such as data availability, or constraints relative to new data capture and processing. Indeed, factors such as acquisition price, time to data acquisition, data accuracy, knowledge to analyze the data, existence of processing solutions etc. are elements that directly impact the type of technologies that will be used to obtain the required information. The lack of resources and knowledge may therefore lead to a sub-optimal pre-selection of data capture and processing solutions, potentially leading to more uncertainties in a less accurate risk picture depiction.

In addition, gaps in tools and knowledge to process the data may also lead to a data accumulation paradox. In such situations, some data types may largely be collected (e.g., inspection images, past outage reports) without further processing. This inefficiency can be costly for utilities. It can lead them to miss opportunities to adequately capture the true risk level, therefore missing opportunities to improve their grid's resilience. There is thus a need for new tools enabling to facilitate the decision-making of the operators.

2.4. The vegetation case-study

Vegetation along power is a well-known and a main source of outages in the field of power grid management (Doostan et al., 2019; Eggum, 2019; Hansen, 2018). Vegetation can be damageable in two ways: (1) by growing under a power line until it connects two phases and thus creates shortcuts, or (2) by falling on the power lines. Although the costs of repairing collapsed infrastructures are relatively acceptable, the disturbances that can result from such events can be quite massive. In fact, tree falls on power lines can result in large-scale wildfires (Doostan et al., 2019; Kumagai et al., 2004). Tree falls have also been reported as main contributing factors to some well-documented blackouts (Haes Alhelou et al., 2019; Sforna and Delfanti, 2006; U.S.-Canada Power System Outage Task Force, 2004).

Vegetation is, by nature, both a simple and challenging hazard. It is simple in the sense that the few scenarios involving vegetation are well-known. Vegetation growth is also a relatively slow process. However, vegetation management is also challenging because the vegetation can grow in most areas where power grids have been installed around the globe. A complete risk depiction regarding vegetation would therefore require a relatively frequent inspection of all power lines where vegetation may grow. This can be particularly costly, increase safety-related risk for the personnel in harsh and remote environments, and is therefore not feasible in practice most of the time.

Partial preventive visual inspections are thus executed to cope with such challenges. Such inspections enable to estimate where future clear-cutting operations will be required. The most common way to handle visual inspections is by sending technicians flying over the lines or walking along them (Matikainen et al., 2016; Nguyen et al., 2018). Alternatively, and/or in parallel, point clouds are also utilized to make more precise distance measurements.

However, the calendar-based approach usually used to plan the power line inspections is known to be sub-optimal. Indeed, it may result in missing the identification of an area where vegetation growth was faster than expected. Additionally, it may result in a sub-optimal use of resources by sending teams to areas with low vegetation growth rates. There is thus a need to improve the way information is managed (i.e., captured, transmitted for analysis, pre-processed, processed, transmitted for reporting to the end-user) with regards to power grid-related risk analysis focusing on vegetation.

3. Challenges Reporting & related Research Questions

This thesis is centered around two main topics:

- Dynamic Risk Analysis,
- Power grid management.

The previous sections introduced those topics and reported different research gaps. The identified research challenges are more precisely characterized in the following sub-sections. This characterization is organized within two blocks. The first of these blocks is dedicated to DRA, while the second is more relative to power grid management.

3.1. Block 1: DRA-related challenges and research questions

3.1.1. Data sourcing

The lack of data is a common challenge for Quantitative Risk Analysis (QRA) (Aven, 2008; Nordgård, 2010) and has logically also been reported as problematic for risk analyses in power grids (Bell et al., 2012; Ciapessoni et al., 2016; Perkin, 2018). There is thus a need to investigate new ways of integrating data into risk models. The research of alternative data sources is also a way to go beyond the consideration of outdated data used for frequency evaluation, which remains a recurrent problem in risk analysis (Creedy, 2011). The digitalization of our society offers new opportunities to address those issues. However, different types of data sources provide different information levels, and each newly considered type of data source's informative potential needs to be investigated in practice.

A first main research question for block 1 can thus be formulated as follows:

Research question 1.1: How can we increase access to data to facilitate the application of QRAs in practice?

This research question can be further divided into two sub-questions:

Research question 1.1.1: Which additional data sources can be considered helpful in risk analysis?

Research question 1.1.2: To which extent does the problem formalization impact the pertinence of data source selection in a risk analysis process?

3.1.2. Conventional risk assessment framework augmentation

The development of a standard is a time and energy-demanding process due to the procedures to be followed and the need for deepened evaluations. Although this provides guarantees regarding the pertinence of the delivered recommendations, it is also a synonym of inertia and reduces the reactivity regarding the acknowledgment of configuration changes. This can also be observed with conventional risk assessment frameworks, which have usually been developed with a mindset of data scarcity and the use of handbook data. Additionally, multiple constraints requiring adaptations of the solution developed in the design phase are only acknowledgeable in operation. Consequently, there is nowadays no efficient solution able to handle multiple heterogeneous datasets dynamically in a real-case big data configuration.

A second main research question for block 1 can thus be formulated as follows:

Research question 1.2: Which improvements shall be provided to conventional risk assessment frameworks to enable dynamic heterogeneous dataset integration in a real-case big data configuration?

This research question can be further divided into two sub-questions:

Research question 1.2.1: Which elements of conventional risk assessment frameworks do require reconsideration or reinforcement in a real-case big data configuration?

Research question 1.2.2: How can the benefits of machine learning be exploited to support better risk analysis?

3.2. Block 2: Power grid-related challenges and research questions

3.2.1. Power grid-dedicated risk-focused solutions

The regular emergence of new data capture solutions provides new opportunities for power grid operators to be better informed of the status of their grid. However, grid operators might have difficulties in understanding the implications of each data capture process on their capacities to optimally manage the grid (Nordgård, 2010). In addition, the issues identified in the field of risk analysis (i.e., lack of methods enabling to capture interactions and dynamic aspects of risk variations (Yang and Haugen, 2015)) also impacts the development of solutions in the field of power grid management (Perkin, 2018). Finally, real-world constraints directly affect the feasibility of solution development. For instance, the success of the creation of a solution is strongly dependent on aspects such as

- the intensity of the demand in the market,
- the available analytical capacities of the involved stakeholders,
- the available software development capacities used to build the product,
- a strong business plan ensuring that the solution is financially viable,
- etc.

Thus, there is a continuous need to develop further and improve risk-based tools usable in practice to optimize the decision-making processes in large-scale power grid management. The solutions to be developed need to ensure that the operator needs are properly addressed, offering different alternatives and highlighting remaining challenges.

A first main research question for block 2 can thus be formulated as follows:

Research question 2.1: How should solutions be designed to enable meaningful comparison and combination of heterogeneous datasets for risk analysis applied in the field of power grid management?

This research question can be further divided into three sub-questions:

Research question 2.1.1: What are the development possibilities and requirements for developing solutions enabling optimized risk analysis in power grids considering heterogenous data sources?

Research question 2.1.2: What are the main challenges for merging various datasets for the purpose of risk analysis?

Research question 2.1.3: Which additional non-analytical constraints shall be considered in the development of a technical solution?

3.2.2. Historical and future data-use optimization

Data management challenges directly impact power operators' capacity to extract all the value of the historical data they have. Additionally, the challenges of understanding the implications of each data capture process might lead to sub-optimal decision-making regarding future data acquisitions. There is thus a need to provide power grid operators with further insights enabling them to unlock the value of their database.

A second main research question for block 2 can thus be formulated as follows:

Research question 2.2: How can we support power grid operators in optimizing the exploitation of their database in the context of risk analysis?

3.2.3. Vegetation management improvement

Inadequate vegetation management along power lines can result in large economical and societal costs. Examples of such impacts are large-scale wildfires (Doostan et al., 2019; Kumagai et al., 2004) and blackouts (Haes Alhelou et al., 2019; Sforna and Delfanti, 2006; U.S.-Canada Power System Outage Task Force, 2004) – among others. Thus, it is necessary to provide power grid operators with solutions enabling them to reduce the risk of vegetation-related outages.

A third main research question for block 2 can thus be formulated as follows:

Research question 2.3: How can new technologies facilitate vegetation management along power lines?

4. Research Objectives

4.1. Definition of objectives

The present thesis's objective is to contribute to both the fields of Dynamic Risk Analysis and power grid management. The following research objectives are therefore defined based on the research question reported in section 3.

4.1.1. Block 1

Research Objective 1.1: Explore, understand & prepare extra data integration possibilities for risk analysis.

- **Sub-objective 1.1.1**: Identify alternative data sources, as well as multiple data capture and processing technics & tools.
- **Sub-objective 1.1.2**: Assess the impact of data resolution variation on the risk analysis performances.

Research Objective 1.2: Define a strategy to augment the data integration capacities of conventional risk assessment frameworks in a real-case big data configuration.

- **Sub-objective 1.2.1**: Identify the adaptation requirements of existing frameworks in a real-case big data configuration.
- **Sub-objective 1.2.2**: Explore the unlocking potential of machine learning for the field of risk analysis.

4.1.2. Block 2

Research Objective 2.1: Support the development of power grid-dedicated risk-focused solutions.

- **Sub-objective 2.1.1**: Assess various construction logics and related data integration policies for risk-focused solutions to be used in the field of power grid management.
- **Sub-objective 2.1.2**: Identify potential stumbling blocks of data integration in the context of risk analysis.
- **Sub-objective 2.1.3**: Appraise real-world non-analytical constraint implications.

Research Objective 2.2: Support value generation from existing data and support decision-making regarding future data captures in the context of power grid management.

Research Objective 2.3: Enable vegetation management improvement.

4.2. Overview of articles and link to research objectives

Figure 4-1 graphically illustrates the linking between the research objectives addressed in the present Ph.D. and reported in sub-section 4.1. Note that the different objectives presented in this thesis were continuously addressed within a cyclic process. As such, most of the steps towards the final results have enabled to address multiple objectives simultaneously. For the sake of simplicity, we distinguish between the content of the two blocks but recall the tight bounds existing between them, as is illustrated via multiple arrows visible in Figure 4-1. Figure 4-1 also details which articles reported in the Declaration of Contributions and present in Part II of the thesis contain response elements for each of the addressed objectives, as is further described below.

The articles II, III, IV, VII, X provide background information enabling understanding the state-of-the-art in risk analysis and support identifying research gaps in that field.

The articles I, II, III, IV, VI, VII, VIII, IX, X provide background information enabling to obtain an adequate business understanding in the field of power grid management and support the identification of stumbling-block hindering the application of risk-related solutions in that field.

Article I addresses the problem of crack formation in Norwegian power grid wooden poles. It suggests investigating the combination of (1) databases reporting automatically identified cracks on assets, with (2) datasets reporting meteorological conditions. It addresses failures on microscopic level.

- ⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1 - 1.1.2; 2.1.2; 2.2.

Article II focuses on the challenges of processing heterogeneous data for risk assessment in a smart grid configuration. For this, it comes back on the role of IT and provides a list of challenges that have been identified in smart-grid configurations. It then uses a bow-tie diagram to illustrate the concept of risk applied to vegetation along power lines. Finally, it provides an extensive list of relevant data sources and suggests combinations possibilities as research avenues.

- ⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1; 2.1.2 - 2.1.3; 2.2; 2.3.

Article III suggests supporting the application of Dynamic Risk Analysis to power grids in real-world situations by using Systems Engineering. It investigates the creation of a dynamic risk analysis pipeline showing how different technologies providing different levels of insights can be combined to address the problem of vegetation along power lines.

- ⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1 - 1.1.2; 1.2.2; 2.1.1 - 2.1.2; 2.2; 2.3.

Article IV focuses on the optimization of technology-based decision-support for managing power girds exposed to the vegetation hazard. It describes this problem using a bow-tie diagram. It then suggests a data-driven approach to define the arrangement of the analytical steps in the processing flow. The considered approach is based on needs expressed by the end-users (i.e., grid operators) and initially ignores any construction constraints by assuming large access to data.

⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1; 2.1.1 - 2.1.2; 2.2; 2.3.

Article V focuses on lessons from past hazardous events and explores the possibilities of using machine learning for severity prediction by analyzing the Major Hazardous Incident Data Service (MHIDAS) database.

⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1; 1.2.2; 2.2.

Article VI focuses on automated power line vegetation monitoring using high-resolution satellite imagery in an industrial context. It first explores how information from LiDAR point clouds can be combined with satellite imagery when addressing the risk induced by vegetation along the lines. It then shows how machine learning can be used to automatically process satellite imagery.

⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1 - 1.1.2; 1.2.2; 2.1.1 - 2.1.2; 2.2; 2.3.

Article VII focuses on developing a method enabling automatic dataset selection for dynamic risk analyses in real-world configurations. The approach starts by reporting reinforcement actions required to adapt existing conventional risk assessment frameworks. A metadata-based method is then suggested to manage datasets originating from heterogenous data sources. Finally, the article uses a case-study with simulated datasets focusing on vegetation along the lines to illustrate its applicability.

⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1 - 1.1.2; 1.2.1 - 1.2.2; 2.1.1 - 2.1.2 - 2.1.3; 2.2; 2.3.

Article VIII is an application to a real-world case-study of the method proposed in article VII. 17 datasets identified as potentially informative for a vegetation-focused power grid risk assessment are considered in the study. The selection of those 17 datasets results from a preliminary analysis enabling to define which type of data sources may provide information usable in a vegetation-focused power grid risk assessment.

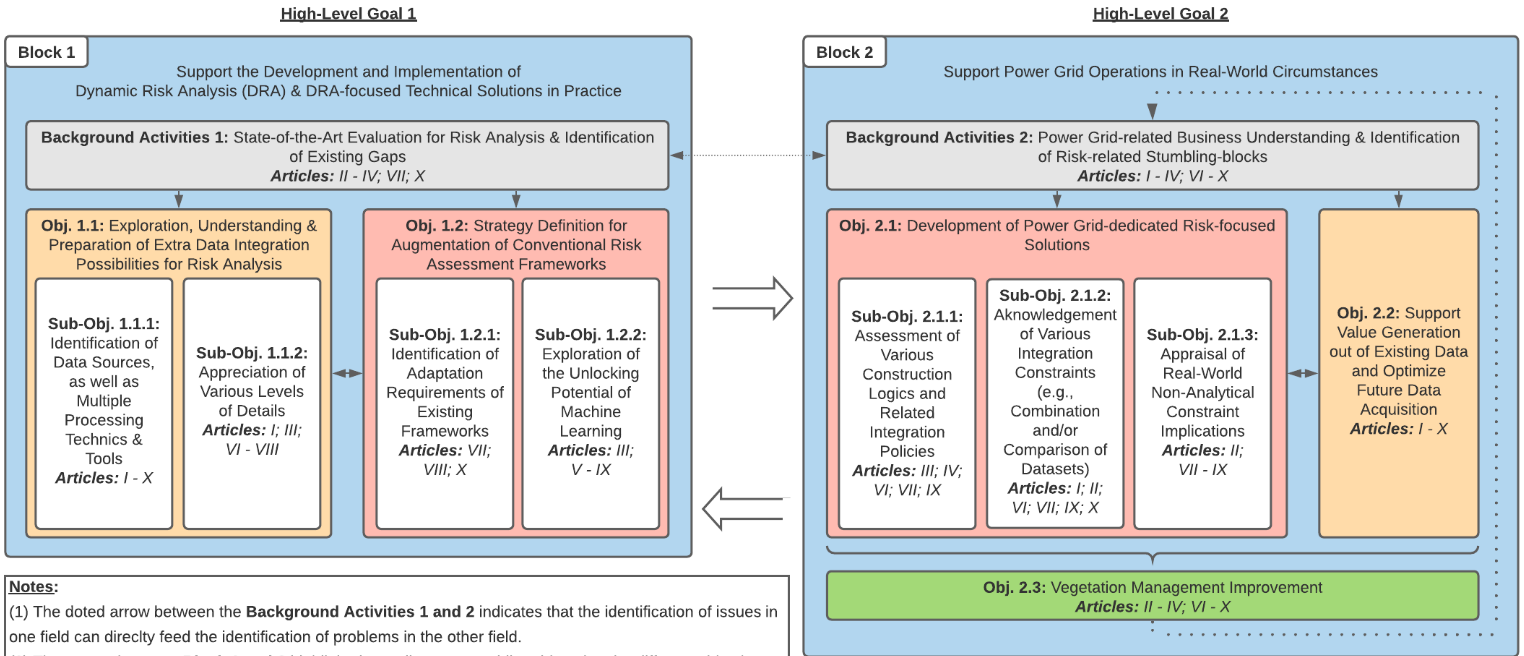
⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1 - 1.1.2; 1.2.1 - 1.2.2; 2.1.3; 2.2; 2.3.

Article IX focuses on developing a platform enabling the merging of heterogeneous datasets in the context of power grid management. The article integrates the analysis of multiple types of data sources for 2 different hazards (wind and vegetation). It uses outcomes from the study reported in Article VI and includes them in an analytical pipeline enabling improved risk analysis in a real-world context.

⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1; 1.2.2; 2.1.1 - 2.1.3; 2.2; 2.3.

Article X addresses the importance of data veracity during data acquisition for risk assessment processes. It suggests building on existing definitions of risk and includes a veracity indicator in the definition to quantify the level of (dis)agreement between different datasets used for the analysis of a specific scene. The indicator is used in a small case-study focusing on datasets informing about the proximity of vegetation in the surrounding of power lines.

⇨ As such, this article is relevant for addressing the following objectives and sub-objectives: 1.1.1; 1.2.1; 2.1.2; 2.2; 2.3.



Notes:
(1) The doted arrow between the **Background Activities 1 and 2** indicates that the identification of issues in one field can direclty feed the identification of problems in the other field.
(2) The arrows between **Block 1 and 2** highlight the cyclic process while addressing the different objectives.
(3) The curly brace in **Block 2** indicates that the obtained insights are used to address the problem of vegetation management.
(4) The doted feedback loop in **Block 2** indicates that the acquired experienced shall continuously be used to increase the understanding of the system. The arrow is dotted to highlight the fact that this feedback only represents a fraction of the experience to consider - the rest corresponding to the experience acquired while addressing other hazards.

Figure 4-1: Graphical reporting of the objectives addressed in the present doctoral project. The articles enabling to address each of the reported objectives are also indicated where relevant.

4.3.Research scope

This thesis mainly focuses on supporting the development and implementation of dynamic risk analysis in an industrial context by optimizing information management upfront such analyses. It does so in the field of power grid management by maximizing the benefits of new non-conventionally used data sources. It also uses both advanced and integrated techniques while simultaneously acknowledging the importance and impact of real-world constraints. Although most of the results may easily be extended to other configurations and used for other risk analyses, it is essential to mention that the initial studies considered the following scope boundaries:

- The work is achieved in a Norwegian environment.
- Only data sources that were relevant for risk analysis in power grid operations were investigated.
- The main type of elements of the grid analyzed were overhead power lines. Although the management of substations has been explored, it didn't consist of the main part of the work. This also implies that power production sites and power consumption nodes fall outside the scope of this research.
- We mainly focused on the hazard "vegetation". Although alternative hazards have been addressed, those weren't the central part of the work.
- Although we indicate how the risk analysis may be implemented in various pipelines, we don't provide specific risk analysis results. This also explains why we limited the investigations relative to risk analysis to the *probability* of trees falling on power lines when developing solutions. This implies that the *consequence* dimension of the risk definition is not part of the main research scope.

5. Research Context & Working Configuration

5.1. Research context

This thesis was prepared within the context of the industrial Ph.D. project named "*Dynamic risk management for Smart Grids in large-scale interconnected power systems.*". Industrial Ph.D. projects consist of close collaborations between a company and a degree-conferring institution. They aim to boost relations between research organizations and the industry. They especially enable increasing the value obtained from academic research by facilitating knowledge transfers to the society and the business sector. They also permit long-term competence-building for the Norwegian industry by recruiting doctoral candidates. For an industrial company, the industrial Ph.D. scheme enables it to reinforce its analytical teams and facilitates its R&D activities. It also allows the acknowledgment of state-of-the-art methods and standards that have been prepared within research and standardization institutions. It is also beneficial for academia, as it facilitates real-world feedback and competence sharing with researchers spending a substantial part of their time in a university environment.

The industrial Ph.D. scheme promotes and enables closer cooperation between all the stakeholders relevant to the project. It allows creating a direct link between the end-users, the researchers responsible for the conception of the solution and the engineers responsible for the development of the product. This ensures a better acknowledgment of real-world constraints, which de facto maximizes the chances of obtaining a final product adequately addressing the user needs originally expressed. Finally, the industrial Ph.D. scheme meets the same academic requirements as the general doctoral degree education and has the same high scientific quality and standard level. It must follow the same regulations for doctoral degree education concerning admissions, implementation, and evaluation as other doctoral projects undertaken at the degree-conferring institution.

The present thesis was funded by eSmart Systems and the Norwegian Research Council (NæringsPh.D. program - Project No.: 276404). The degree-conferring institution ensuring that the doctoral project maintained an adequately high scientific standard is the Norwegian University of Science and Technology (NTNU - Norway). The supervision of the thesis on the side of NTNU was ensured by Nicola Paltrinieri and Pierluigi Salvo Rossi. Nicola Paltrinieri is a full professor of risk analysis at NTNU and an adjunct professor in offshore HSE management at the University of Bologna (Italy). He has earned a Ph.D. in Environmental, Safety and Chemical Engineering from the University of Bologna. Pierluigi Salvo Rossi is currently a full professor of statistical machine learning with the Department of Electronic Systems at NTNU. He received the Dr.Eng. degree in telecommunications engineering and the Ph.D. degree in computer engineering from the University of Naples ``Federico II" (Italy). The supervision of the thesis was also ensured by Davide Roverso, Chief Analytics Officer at eSmart Systems. Davide Roverso holds a Ph.D. degree in Computing Science. He has over 30 years of experience in the field of Machine Learning and Big Data Analytics, with applications in diagnostics, prognostics, condition monitoring, and early fault detection in complex processes, in sectors ranging from energy to medicine and environmental monitoring.

5.2. Activities

All the academic activities required for the preparation and validation of a doctoral project taking place at NTNU were fulfilled during this thesis. This implies:

- validation of Ph.D. courses,
- literature reviews for acknowledgment of the required information in the different relevant fields,
- preparation of multiple publications (journal articles and conference articles) anonymously reviewed by relevant experts,
- participation in multiple international conferences.

In addition to this academic basis, the research achieved during this Ph.D. also strongly relies on an involvement in multiple projects. Their ambition and my implication in those projects are shortly described in the following. The chronology of all the activities is illustrated in Figure 5-1.

5.2.1. Project 1 – The SAIRA Challenge

France operates the largest power distribution network in Europe (almost 1.4 million kilometres of power lines, around 800 000 transformers, and 37 million customers) (Enedis, 2020, 2017). Enedis represents one of the main actors in this context. It manages 95% of continental France's electricity distribution network and is responsible for development, operation, asset maintenance, and continuous public electricity service. In order to accelerate defect identification, increase detection accuracy, reduce costs and verify the usability of Unmanned Aerial Vehicles (UAVs) as a platform for collecting inspection data, Enedis launched in April 2018 an "Automated Image Analysis System" challenge (or SAIRA Challenge, for "Système d'analyse d'images réseau automatisé"). eSmart Systems was one of four leading companies with core expertise in the field and in computer vision, chosen to participate in the challenge. The project took place from January 2019 to June 2019. eSmart's solution ranked as the #1 performing system for defect detection, both in terms of accuracy and speed.

Involvement:

I handled the application for eSmart Systems, as well as the contract negotiations before starting the project. I then managed the project and facilitated the technical and administrative communication between the different stakeholders. Finally, I also executed a quality assessment on the results that we provided and handled the delivery to Enedis. The project especially enabled me to understand Enedis' needs and to acquire more experience in the field of computer vision applied to power grids.

5.2.2. Project 2 – Astrum

Astrum was an R&D project lead by eSmart Systems and executed in collaboration with 5 Norwegian DSOs. The main objective was to create an intelligent decision support tool to facilitate access to relevant information in the context of power grid management. That way, grid operators can better plan, develop and maintain the network. The final ambition was to optimize the management of the grid operations and facilitate meaningful reinvestments in an aging infrastructure. The project was successfully closed at the beginning of January 2020.

Involvement:

I participated in multiple workshops with the different DSOs. Those exchanges first enabled me to better grasp the nature of the user-needs. They also enabled me to share insights on how power grid operators could generate new data and better use the data they had. Eventually, I proposed a prototype of a solution that could be used to prioritize the visits of substations for maintenance operations.

5.2.3. Project 3 – GridEyeS

The objective of GridEyeS is to support power grid operators with the control, maintenance and planning of tasks relative to the infrastructure they manage. It aims to do so by combining satellite images, weather information and AI and integrating those in tools enabling grid operators to take decisions. This project aims to enable automatic, cheaper and more accurate detection of risky areas. Based on those insights, grid operators can move from calendar-based maintenance planning to risk-based planning, enabling them to optimize their resources and the performances of their grid. GridEyeS started as a Feasibility Study controlled by the European Space Agency (ESA). The application for the project was initiated in early 2019 and the project was kicked-off in September 2019. The feasibility study was successfully closed in September 2020. We are currently preparing an application enabling to convert the feasibility study into a demonstration project. A demonstration project consists in a 2-year project with the goal of developing a final commercial product. The current plan – subject to validation by the European Space Agency – is to initiate this new phase by the end of 2022. Note that multiple preliminary activities relative to satellite data information were executed within 2018. Such activities mainly consisted in identifying relevant data sources and tools and getting more familiar with satellite image analysis.

Involvement:

I handled the application for eSmart Systems. I then managed the project on the side of eSmart Systems and handled the communication with the power grid company involved in the project. I took part in all the discussions with the European Space Agency. I strongly contributed to both the conception of the platform hosting the analytics, as well as the analytics themselves. Finally, I executed a quality assessment of our results and co-handled the delivery of the deliverables required by ESA.

5.2.4. Project 4 – Connected Drone 2 & Project 5 – SkogRiskAI

Connected Drone 2 (CD2) is an R&D project lead by eSmart Systems and executed in collaboration with 22 DSOs. It comes as a direct follow-up to the first Connected Drone project, also lead by eSmart Systems. CD2 was kicked-off at the end of 2018 (the same day CD1 was closed) and was closed at the end of 2021. CD1 was mainly focused on developing an innovative product that automatically detects and displays assets and component defects on aerial power line inspection images. CD2 went several steps further. First, it aimed at reinforcing the product developed so far. Second, it also aimed at providing several additional intelligent features. For instance, autonomous flights and use of simulated environments are two of the main research avenues for CD2.

SkogRiskAI is an R&D project lead by eSmart Systems and executed in collaboration with NIBIO (the Norwegian Institute of Bioeconomy Research) and 5 Norwegian DSOs. The main objective of the SkogRiskAI project is to develop an operational solution for better risk-informed decision-making concerning vegetation along the power lines. This is influenced by various parameters, such as local vegetation properties, data collection method, etc. The purpose of the solution to be developed is in particular to facilitate the prioritizing of clear-cutting operations as well as to be better prepared in case of emergencies. The project brings together computer science experts, risk analysts, forestry experts and grid operators. It explores different sensing technologies to obtain an optimal cost-benefit ratio depending on the size of the grid operator. Costs and benefits can, for instance, be assessed concerning data acquisition time, data capture price, data processing time, precision and accuracy of the results – among others.

Involvement:

I acquired experience with 3D point clouds in the first half of 2018 (preliminary activities). This experience mainly consisted in identifying tools and technics to generate and analyse point clouds (LiDAR-based and photogrammetry-based). I shared my experience regarding 3D point clouds for the project CD2, especially because of its relevance for a digital representation of an infrastructure. Similar types of insights have been provided for the SkogRiskAI project. In that last case, I also advised and shared knowledge regarding vegetation management, data analysis, risk analysis and data transformation technics.

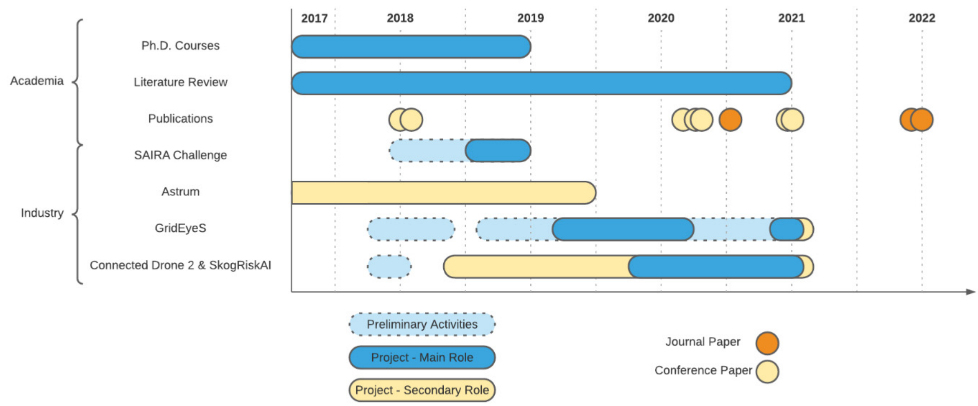


Figure 5-1: Distribution of the academical and industrial activities executed during the doctoral project.

6. Research Approach

6.1. Fundamental characterization of the research process

The industrial Ph.D. scheme only supports projects that can be classified as "industrial research" and/or "fundamental research" (Research Council of Norway, 2021), as is defined by the European Commission (European Commission, 2014b). Both definitions are provided in (European Commission, 2014c) and reported as following:

Definition Industrial Research: "Industrial Research means the **planned research or critical investigation** aimed at the acquisition of **new knowledge and skills** for **developing new products, processes or services or for bringing about a significant improvement in existing products, processes or services.** It comprises **the creation of components parts** of complex systems, and may include the **construction of prototypes** in a laboratory environment or in an environment with simulated interfaces to existing systems as well as of pilot lines, when necessary for the industrial research and notably for **generic technology validation**".

Definition Fundamental Research: "Fundamental Research means **experimental or theoretical work** undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts, **without any direct commercial application or use in view**"

The research objectives I address in this thesis (sub-section 4.1) as well as the research configuration of this project confirm that the present Ph.D. can be classified as Industrial research. Indeed,

- The **definition, coherence, and systematic execution of the research plan** of the current doctoral project was ensured by eSmart Systems, the Norwegian University of Science and Technology and the Norwegian Research Council. The work can also be considered as a **critical investigation** considering the number and the variety of experts involved at each project step. The multiple publications of this work were also **peer-reviewed by external experts and finally validated by the end-users** (i.e., grid operators), who can be considered as the final reviewer of the results as subject-matter experts.

- We are addressing the research questions reported in section 3. Those are based on **previously identified knowledge gaps and limitations currently faced in practice**. We address those research questions by considering both **new technologies** and **latest advances** in multiple research fields.

- The willingness for eSmart Systems to reinforce and **construct new products** is obvious as a commercial company. The company needs to continuously invest in research and innovation **to improve how user-needs are addressed** nowadays. The present thesis contributes to that ambition in multiple ways as is reported in section 7.

The current work also meets additional criteria required to characterize a work as an R&D project (OECD, 2015):

- Creativity: the project mainly builds on a **Systems Engineering mindset to break silos** and **combines multiple fields of expertise** in a meaningful and innovative way.

- Uncertainty: the final outcome of the project was uncertain, especially because we used to be unsure how multiple heterogeneous datasets could be combined in a **systematic way**.

- Transferability and/or reproducibility: the generic aspect (application-field agnostic) of the objectives reported for block 1 in sub-section 4.1 shows how **this Ph.D. also contributes to more *Fundamental Research***, facilitating that way a **transferability of the results**. Additionally, **special attention** has been given to reporting all the assumptions we made, the tools we considered, and the data sources we used in the present work. This **facilitates the reproducibility** of the work.

The quality of the research executed in this thesis was ensured via different mechanisms. In particular, all publications reported in the present thesis are:

- based on objectively identifiable references and standards; they are also referring to the work of multiple recognized institutions;
- co-written by authors with internationally recognized expertise;
- anonymously peer-reviewed by experts in the relevant fields through the different journal and conference reviewing systems;
- prepared in close collaboration with multiple subject-matter experts (Power grid operators, risk analysts, satellite image analysts, vegetation-related researchers, etc.), enabling thus to benefit from both direct inputs and feedback;
- using real-world or real-world-like data usable within solutions to be developed.

Finally, the experience and the expertise of the supervisors of this thesis enabled a constant guidance and verification of the work achieved during the Ph.D.. This enabled to keep track of each step completed, ensured consistency of the work, as well as a critical review to assure the significance of the provided results.

6.2.Research strategy

A system can be defined as a set of interconnected elements, which aggregation is engineered to meaningfully realize or achieve a specific function or objective (Meadows, 2009). Systems Engineering (SE) can be defined as the interdisciplinary approach aiming at successfully developing a system (INCOSE, 2021a). It strongly advocates for three main principles:

- Adequately understanding, considering and documenting the customer needs all along the system's life cycle, starting early in the development phase.
- Enabling circularity in the thinking/acting process.
- Providing both:
    - a detailed picture of the system’s elements to understand their individual functioning, and

- providing a full picture of the system to understand the interactions between its elements.

SE enables to systematically and adequately structure the research. Its cyclic aspect enables to dynamically acknowledge any adaptation requirement throughout the life cycle of a system and maximizes the probability of encompassing all relevant stakeholders' points of view.

Systems engineering performance strongly depends on the acknowledgement of existing tools, techniques, methods, experience, standards, principles, and concepts (INCOSE, 2021b). The larger this background knowledge, the higher the chances of finding initial solutions and potential alternatives to problems reported by the different stakeholders. The final performance evaluation enables to characterize how well the developed systems support the decision-making of the end-user, that way defining the level of satisfaction of the customers (Haskins, 2008).

Therefore, the present thesis relies on using a SE mindset to address the different objectives reported in sub-section 4.1.

Additional task-specific methods have also been used for the realization of the present thesis. Those can be grouped in the following clusters:

- **Literature review & business understanding.**
  The fields that were explored can be listed as following: risk analysis, machine learning, metadata analysis, meta-learning, data quality analysis, power grid functioning, power grid management.

- **Technological and technical benchmark for data acquisition and processing possibilities.**
  The main data sources that were explored can be listed as following: satellite images, aerial orthophotos, aerial inspection images, photogrammetry point clouds, LiDAR point clouds, weather records, multiple GIS mapping systems. An extensive list of sources for which data is publicly available has been reported in appendix B of Article VIII. The tools used during this Ph.D. to analyse such datasets are available in Table A of Appendix A.

- **Knowledge sharing.**
  Most of the contributions provided in the present thesis came as a result of diverse collaborations. Those consisted of workshops, group discussions, brainstorming activities and communications with experts in various fields. Such close and regular contacts with the different industrial partners and other relevant stakeholders was particularly facilitated in the industrial Ph.D. configuration.

6.3.Research methods

A part of the relevant research background referring to the fields and methods mentioned in the research strategy has been provided in section 2. The present subsection succinctly reports additional details relative to the most relevant and potentially less trivial methods not yet described. Those methods were used as tools to identify and process the various datasets relevant to the thesis. The reader is directed to the articles in Part II of the thesis for further details on their application.

6.3.1. Bow-tie analysis (Articles: II, IV)

The bow-tie analysis is a well-established method (Paltrinieri and Khan, 2016). It highlights the relations between a critical event (CE), the scenarios that can lead to the CE (causes), the scenarios that can result from the CE (consequences) and the barriers enabling either to prevent the causes or mitigate the consequences of the critical event (Delvosalle et al., 2006). As such, the bow-tie diagram can be seen as the combination of both a fault tree (FT) and an Event Tree (ET). An FT follows a top-down logic and enables to report the link between the CE and single basic events (e.g., component failures, environmental conditions, human errors, and normal events). On the other hand, an ET follows a forward logic. It enables to map the different paths that can be followed after the occurrence of a CE. Both the Fault Tree and the Event Tree can be used as a qualitative and/or quantitative. Therefore, the bow-tie diagram can also be used as in a qualitative or quantitative way (Raussand, 2011).

A bow-tie diagram (see Figure 6-1 as an illustration) is thus an easily understandable tool that can particularly facilitate the reporting of risks. Note that multiple critical events can result from a unique hazard and that one bow-tie diagram should be prepared for each of the critical events.

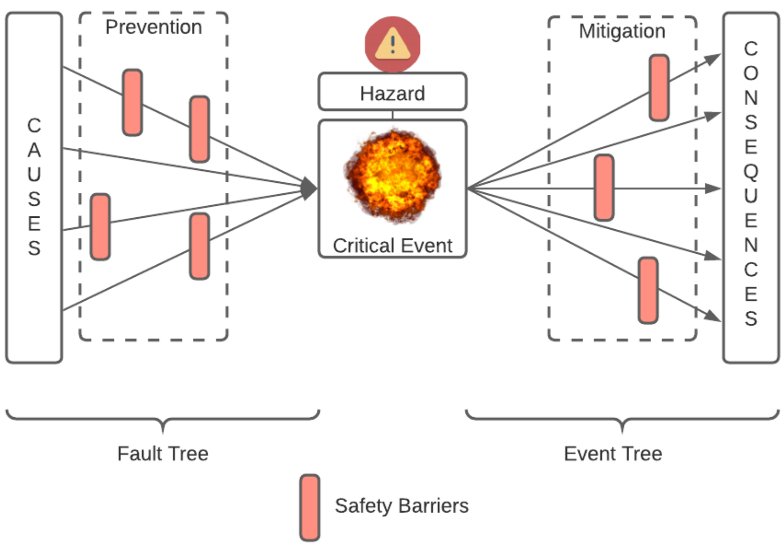


Figure 6-1: General representation of a bow-tie diagram.

6.3.2. Machine learning (Articles: III; V-IX)

Machine learning is a broad and well-known field, which aims to teach computers how to learn from experience (Samuel, 1959). Machine learning-based solutions enable to automate redundant processes, therefore providing numerous opportunities to save resources. They make use of recognizable patterns to facilitate the decision-making of the end-user. Methods used in the early ages of machine learning were mainly statistic-based. For instance, the multiple linear regression (MLR) is one of the simplest statistical technics. It uses a vector of inputs (i.e., the explanatory variables) to predict the value of a response variable (Andrews, 1974). An input vector $X^T = (X_1, X_2, ..., X_p)$ and the output Y can thus be linked by using equation (Eq. 6-1) (Hastie et al., 2009):

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j \tag{Eq. 6-1}$$

where $\hat{\beta}_0$ is the defined as the bias and the coefficients $\hat{\beta}_j$ represent the model weights. The coefficients are estimated using a training set and the model is eventually used to generate outcome predictions for a new unseen input vector.

Although machine learning is not a new field, it has regained more attention in the last decade for 3 main reasons:

- **Data generation and storing capacities.**
  More and more data are made available by the digitalization of our society. Additionally, storing costs have drastically decreased. A white paper from the International Data Corporation (IDC) predicted in 2018 that the global datasphere would reach up to 175 Zettabytes by 2025 (Reinsel et al., 2018). This number can nowadays be expected to be revised upwards due to the COVID-19 pandemic and the extensive use of remote working.

- **Deep Neural Networks (DNNs).**
  DNNs correspond to Artificial Neural Networks (ANN) with a multitude of hidden layers between the input layer and the output layer. The concept of ANNs (i.e., having a computer program trying to mimic the brain functioning (Goodfellow et al., 2016)) is a relatively ancient concept as well (Rosenblatt, 1958). However, DNNs models are able to integrate multiple levels of abstraction. They do so by dividing a complex representation (e.g., a car) into a combination of simpler elements (e.g., wheels, doors, windows, etc.). Those can then be further divided in even simpler concepts, such as contours, corners, etc. – and so on down to pixel level (Nguyen, 2019). DNNs have proven to be particularly helpful in multiple fields (Computer vision, speech recognition, fraud detection, etc.) but are also particularly resource demanding.

- **Graphics Processing Units (GPUs).**
  The rise of DNNs is directly linked to the possibility of using GPU-accelerated computing (Mittal and Vaishay, 2019). GPUs correspond to specialized processors initially used for memory-intensive graphical computation, such as texture mapping and rendering polygons. This capacity to handle memory-intensive tasks has revealed to be particularly helpful in the development of DNNs and GPUs have enabled to drastically reduce the training and testing times (Mittal and Vaishay, 2019).

Machine learning models can be particularly powerful once fully trained. They have been applied in a diversity of cases (object recognition, self-driving cars, etc.) and have even outperformed humans on tasks such as medical diagnostics, gaming or fraud detection (Brynjolfsson and Mitchell, 2017; Buetti-Dinh et al., 2019; Froomkin et al., 2019). However, most of those models are only performant for a specific type of tasks and can hardly be transferrable to other duties.

Meta-learning comes as a potential facilitator when facing such a problem. It aims to "learn how to learn" by considering both positive and negative experiences during the execution of different machine-learning algorithm (Giraud-carrier, 2008). This experience is then taken into consideration to inform meta-features, which are used to characterize performance differences between the considered algorithms (Vanschoren, 2018). Correspondences between a previously considered task and a newly considered task can then be used to select the models that would most likely provide the best results in the new configuration (Pimentel and de Carvalho, 2020).

6.3.3. Pre-processing actions and use of newly generated features (Articles: I, III, VI-X)

Multiple pre-processing actions were executed after the technological benchmark and the acquisition of the datasets considered for the present thesis. Examples of such pre-processing steps are:

- photogrammetry point cloud generation,
- photogrammetry and LiDAR point cloud filtering and classification,
- various types of distance measurement in the point clouds between pre-selected clusters (e.g., vegetation-power line),
- various projections and mappings,
- aggregation of satellite images,
- satellite images analysis and classification,
- etc.

Exploring the selected datasets and applying those pre-processing steps enabled to identify the parameters that could be informed by the datasets originating from data sources identified as relevant for a risk analysis. This came in addition to the knowledge acquired through discussions with the relevant stakeholders and through the literature reviews.

The globally cumulated knowledge enabled to propose an amelioration of existing risk assessment frameworks. This improvement is first based on reinforcement actions able to (1) adequately characterize the type of information one is looking for in a risk analysis and (2) correctly identify where this information may be obtained. In addition, the improvement is based on the development a metadata-based method (called the *Three-Phases Method*) enabling to reduce uncertainties relative to information acquisition. This method identifies the informative potential of heterogeneous data sources and dynamically ranks the datasets originating from those sources to select the most informative dataset for a risk analysis. A shortened overview of the method is provided as a main contribution of this Ph.D. in sub-section 7.2, while a full description of the method is provided in Article VII.

7. Contributions

The present section reports the main research findings of this doctoral project by briefly summarizing the principal contributions of the articles present in Part II. The contributions in sub-sections 7.1 and 7.2 are mainly focused on the development and implementation of Dynamic Risk Analysis (DRA) & DRA-focused technical solutions in practice, while the contributions in sub-sections 7.3, 7.4 and 7.5 are principally focused on supporting power grid operations in real-world circumstances.

7.1.Contribution I: Increasing data acquisition opportunities for QRAs.

- This contribution aligns with objective 1.1, including sub-objectives 1.1.1 and 1.1.2.
- The following articles contributed to address sub-objective 1.1.1: I, II, III, IV, V, VI, VII, VIII, IX, X.
- The following articles contributed to address sub-objective 1.1.2: I, III, VI, VII, VIII.

The known data scarcity in the field of risk analysis calls for alternative, conventionally and/or non-conventionally used data sources to better picture the status of industrial facilities. Multiple research activities have been executed in that direction and enabled to identify various data capture and processing technics & tools (sub-objective 1.1.1).

All the articles prepared during this doctoral project refer to data sources and datasets providing valuable information for risk analysis in power grid management. The ones I mention in the following enable to cover the complete list of data sources we have identified. Article I looks into a database of automatically classified power poles and suggests a combination with meteorological data to identify links between cracks in wooden poles and weather conditions. Articles II then addresses the problem from a more general perspective and reports a list of data sources that can be used in predefined use-cases. Those use-cases highlight hazards, issues and challenges that utilities can face while managing their power grids. The ways suggested to exploit the reported data sources also represent as many research avenues addressable by the grid operators based on the topics they consider being the most relevant. Article III suggests considering additional data sources such as topographical wind exposure or human population density to better characterize both the probability and consequences of power outages. Article IV additionally proposes as an opening to investigate sources of information enabling to characterize the impact of human factors. Indeed, both internal factors (e.g. distraction, fatigue) and external and organizational factors (e.g. human-system interface quality, procedures, workplace adequacy) influence the operator's decision and need thus to be investigated to reduce the impact of human errors in the decision making process. Article V focuses on the Major Hazard Incident Database (MHIDAS) (AEA technology - Major hazards assessment unit, 2003) to predict the severity of incidents based on historical data. This work indicates thus how power grid operators may use the incident databases they can access but do not necessarily use for quantitative risk analysis.

All articles except I and V reported in Part II focus on a case-study addressing the problem of vegetation along power lines. Article VIII provides the most detailed overview of the type of data sources that we have identified as relevant when assessing the probability of a tree fall on a power line. This list can be reported as follows:

- Orthophotos based on satellite images,
- Orthophotos based on aerial images,
- Aerial optical inspection images,
- Photogrammetry point clouds,
- Meshed photogrammetry point clouds,
- LiDAR point clouds,
- Forest surveys (Map),
- Pests/fungi surveys (Map),
- Soil surveys (Map),
- Topography (Map),
- Topographical wind exposure (Map),
- Weather historical data,
- Weather predictions,
- Clear-cutting operations history,
- Planned clear-cutting operations.

Both Article II and VIII provide links to downloadable datasets, which facilitates access to data to other stakeholders (researchers, grid operators, etc.) and favours reproducibility of the results provided in this thesis.

Once identified and downloaded, knowledge and tools are required to explore the informative potential of the datasets. Table A in Appendix A provides a summary of all the tools that have been used for the realization of the present thesis.

Attempting to include new types of data sources in a risk analysis implies nevertheless to understand how a heterogeneity in the dataset attributes can influence the process. It requires specially to assess the impact of data resolution variation on the risk analysis performances (sub-objective 1.1.2).

Articles I and VI properly illustrate the scale variability that can be encountered when considering the analysis level of a research project. Article I considers an asset database reporting the impact of a phenomenon starting on microscopic level (i.e., apparition of crack in wooden poles due to the hygroscopic behavior of wood (i.e. tendency to absorb humidity)). On the other hand, Article VI focuses on using satellite images to assess the status of vegetation along the lines (macroscopic perspective). Such a scale variability plays an important role in the feasibility of the analysis, both analytical performances and processing costs being directly correlated. There is thus a need to adequately identify how the use of resources shall be optimized when attempting to obtain both a local and global understanding of a large-scale infrastructure.

The work in Article III highlights the acknowledgment of this reality. It proposes to capitalize on the advantages of both levels of analysis. It merges a top-down analysis (macro) with a bottom-up analysis (micro) to optimize resource exploitation in the management of vegetation along power lines. Such a requirement is formulated more generically in the development of the approach and method proposed in Article VII and applied in Article VIII (see contribution II).

7.2. Contribution II: Augmentation of conventional risk assessment frameworks enabling dynamic data integration.

- This contribution aligns with objective 1.2, including sub-objectives 1.2.1 and 1.2.2.
- The following articles contributed to address sub-objective 1.2.1: VII, VIII, X.
- The following articles contributed to address sub-objective 1.2.2: III, V, VI, VII, VIII, IX.

The possibility of considering new, different and more datasets supports the implementation of dynamic risk analyses in real-case configurations. However, this also comes at the costs of some structural changes in conventional risk assessment frameworks and requires the implementation of multiple safeguards (sub-objective 1.2.1).

As a simple initial option, the work executed in Article X suggests looking, prior to the risk analysis calculations, at the level of agreement between heterogeneous datasets informing about the risk level in a specific situation. This cross-validation suggestion builds on an extension of Kaplan & Garrick's definition of risk previously provided by (Aven and Krohn, 2014). Aven and Krohn do indeed suggest including the *knowledge* dimension to the original triplet *scenario*, *probability*, *consequence*. We propose further decomposing the *knowledge* dimension into two indicators: (1) the number of sources available to inform a specific observed parameter and (2) a veracity indicator, capturing the agreement level across the considered data sources. Considering this type of indicators is a common approach for data validation and reconciliation (Narasimhan and Jordache, 2000) to qualify the veracity of information, which motived our suggestion for the field of risk analysis. Integrating this type of quality controls increases the pertinence of results provided by risk analysts by reducing the uncertainty level relative to the validity of the considered data. This becomes even more relevant when potentially considering data sources with lower trust levels.

The work in Article VII provides a complementary, more holistic and more detailed description of the modifications that can be suggested to favor the integration of heterogeneous datasets in conventional risk assessment frameworks. In fact, the benefits and limitations relative to the use of datasets with various resolution level (cf. contribution I) can first be used to highlight the importance of adequately establishing the context prior to a risk analysis. Risk analyses applied in real-world configurations will always be restricted by resource constraints limiting the possibilities of capturing and transmitting information, therefore directly impacting the analytical outcome possibilities. The simultaneous definition of both the objectives of the risk analysis and the resource budget that will be allocated to that task thus directly determine the best level of information that will be reachable for the project. This assessment of the best available level of information (also called "level of convergence") is a task that is suggested as a first reinforcement action within the approach suggested in Article VII.

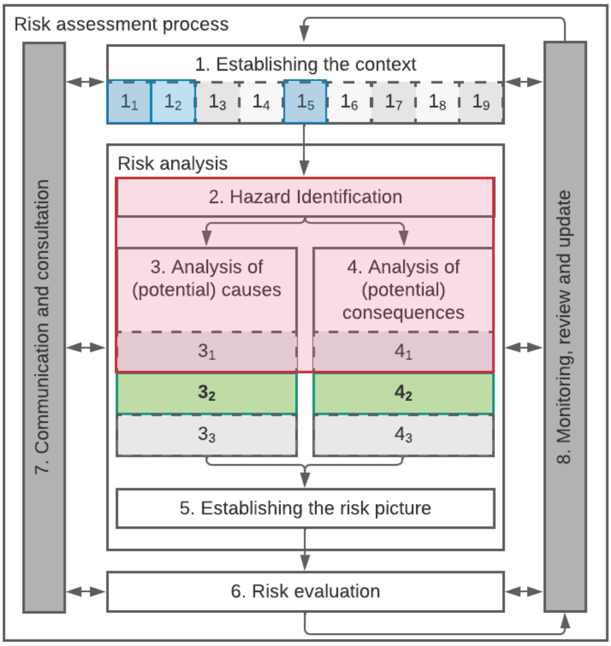
An item of interest on which a risk analysis will be applied needs to be selected once the level of information is clarified. The first series of actions in the risk analysis consists then to:

- apply a hazard identification,
- report all the scenarios for each critical event relative to each identified hazard,

- identify all the dimensions involved in reported scenarios (for both probabilities and consequences),
- identify all the parameters playing a role in the reported dimensions.

The identified parameters are then characterized in terms of unit, resolution, and range. This enables to determine reference values to be used to assess the informative potential of datasets considered in the risk analysis. This characterization is also done from a spatiotemporal perspective considering that risk analyses are time- and space-specific. As a final step, all data sources able to inform the reported parameters to any extent possible are also identified. This strategy is characterized as the second reinforcement action within the approach suggested in Article VII.

Once the reinforcement actions are applied, we suggest further augmenting conventional risk assessment frameworks by integrating a new action within the process with the purpose to dynamically manage heterogeneous datasets. This management shall be done based on the effective informative potential (or "potential of knowledge") of the datasets, which can be calculated using the *Three-Phases Method* proposed in Article VII. Figure 7-1 builds on Figure 2-2 and shows where both the reinforcement actions and the newly suggested actions are applied. Reinforcement action 1 is further detailed in Figure 7-2 and reinforcement action 2 is further detailed in Figure 7-3.



Detailing Step 1 and placement of Reinforcement Action 1 (RA1):

Step 1: Definition of

1_1) Objectives (RA1)	1_2) Scope (RA1)
1_3) Responsibilities	1_4) Methods, models & tools
1_5) System boundaries & System basis (RA1)	1_6) Risk acceptance criteria
1_7) Deliveries	1_8) Execution plan
1_9) Any other relevant activity	

Detailing Steps 2, 3 & 4, placement of Reinforcement Action 2 (RA2) and intergation of new sub-steps for optimized dataset management:

Step 2

Hazard Identification (RA2)

Step 3

Sub-step 3_1: Identification of initiating event causes (RA2)
Sub-step 3_2: Dynamic ranking of informative datasets (Causes)
Sub-step 3_3: Assessment of initiating event frequencies

Step 4

Sub-step 4_1: Identification of initiating event outcomes (RA2)
Sub-step 4_2: Dynamic ranking of informative datasets (Outcomes)
Sub-step 4_3: Assessment of Event sequences, Performances of satefy barriers & Magnitude/extent of the event

Figure 7-1: Augmented risk assessment framework Z-013 (NORSOK, 2010). The steps highlighted in blue and red are the subject of augmentation (reinforcement actions 1 and 2), and the steps 3_2 and 4_2 in green are additional steps related to the optimization of data source/dataset management (Pacevicius et al., 2022).

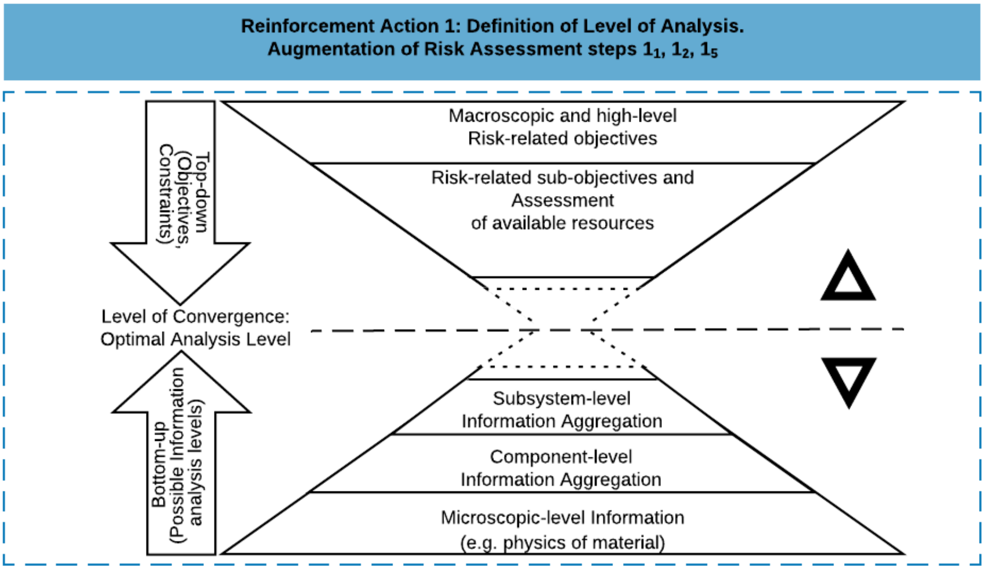


Figure 7-2: Reinforcement Action 1 - Level of analysis of a risk assessment defined as tradeoff decision between stakeholder expectations and analytical possibilities (Pacevicius et al., 2022).

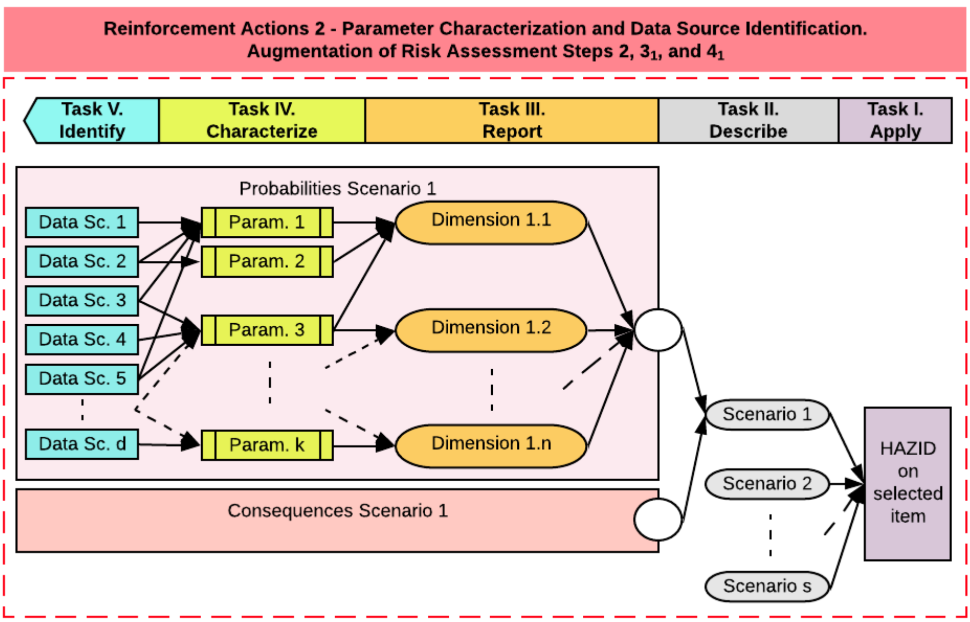


Figure 7-3: Reinforcement Action 2 - Identification of parameter characterization requirements and data sources potentially exploitable for the risk analysis (Pacevicius et al., 2022).

The management of various heterogeneous datasets considered for risk analyses has enabled to learn 3 main lessons:

1) The best level of information a dataset can provide is directly determined by the nature of the file (i.e., its type and format).

2) The informative potential of a dataset considered for a risk analysis is directly impacted by discrepancies existing between the spatiotemporal characteristics of the scene to analyse and the spatiotemporal characteristics of the considered dataset.

3) The informative potential of a dataset considered for a risk analysis is directly impacted by the trust one can assign to the origin of the information.

Therefore, the *Three-Phases Method* aims at building on this knowledge to rank heterogeneous datasets based on the effective potential of knowledge they are synonym of. It capitalizes on the fact that metadata coming with most datasets usually contains information usable to draw conclusions regarding this knowledge. The Three-Phases method, illustrated on high-level in Figure 7-4, characterizes the final informative potential of a dataset by:

1) First, calculating the Default Maximum Potential of Knowledge of the dataset (application of lesson 1).
2) Then, applying a first degradation factor characterizing spatiotemporal discrepancies to obtain a first updated version of the dataset's potential of knowledge (application of lesson 2).
3) Finally, applying a second degradation factor characterizing the trust in the origin of the data to obtain a second updated version of the dataset's potential of knowledge (application of lesson 3).

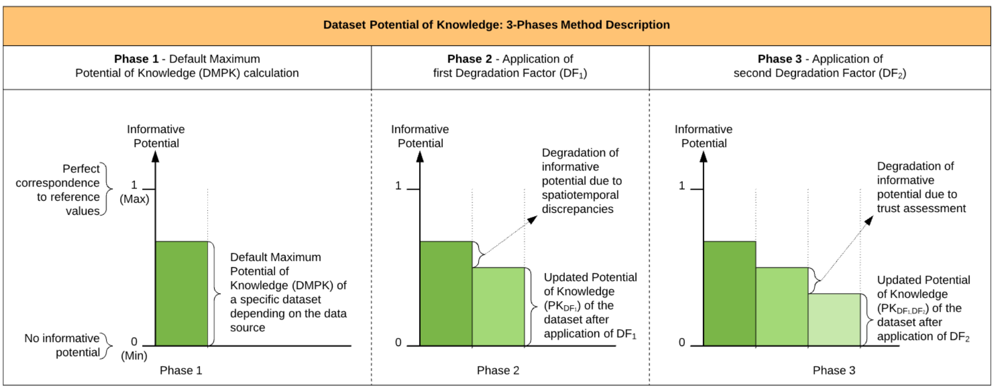


Figure 7-4: Three-Phases Method Description. The figure shows the progressive degradation in the assessment of a dataset's informative potential when compared to the originally required level of information, as defined during the application of reinforcement action 2 (Pacevicius et al., 2022).

Eventually, the metadata-based approach developed in Article VII has multiple advantages. First, it enables to better foresee which level of insights can be obtained based on the definition of the previously reported convergence level. Second, it enables to objectively and quantitatively determine the type of information that must ideally be sought. Third, it highlights when, for which reason and how old and new data sources shall be integrated in the risk analysis process. Forth and foremost, the approach enables filtering large databases more efficiently, optimizing thus the dataset selection process. It keeps only the dataset that minimizes the uncertainties relative to data acquisition and contributes that way to a more accurate risk depiction. As datasets are compared based on reference values relative to the parameters and defined during ideal information quantification, it offers the possibility to assess the informative value of a dataset on a parameter-by-parameter evaluation rather than as a whole, which may otherwise lead to inappropriate discarding of usable datasets. It contributes thus also to making the risk assessments even more dynamic by reducing the data processing time and costs. Finally, the strong industrial context in which the method was developed also maximizes the chance of applicability in real-world situations. This statement is also strengthened within Article VIII, corresponding in an application of the proposed approach on a case-study with real-world data.

The metadata-based approach illustrates how experience can be gained from the machine learning field to support the application of dynamic risk analysis in practice. The essence of dynamic risk analysis is indeed to automatically adapt the risk analysis based on context modification. Context modification may include system structural changes, environment context changes, modelling upgrades or changes in data delivery modes. Machine learning can contribute to the automatic acknowledgment of such changes, favouring that way the implementation of more dynamic risk analysis solution. However, there is nowadays still a limited understanding on how this could be further applied in practice. The exploitation of machine learning-techniques was thus also investigated in alternative ways in this thesis (sub-objective 1.2.2).

A benchmark and analysis of existing applications in the field of machine learning enabled to first conceptually assess how machine learning could support the development of analytical pipeline focusing on risk analysis in power grids (Article III). Based on those conclusions, we decided to develop a computer vision-based solution able to automatically classify vegetation along power lines on satellite images (Article VI). Focusing on a Pléiades-1 satellite image, the final solution enabled for instance to correctly detect areas with low outage probability by 98.2% and areas with high outage probability by 84.6%. Both the outcomes of Article III and VI eventually inspired the GridEyeS platform developed and presented in Article IX.

Finally, machine learning was also used for preparing Article V, where we explored how one could automatically learn from past experiences to improve future risk management. The idea was to train a computer to predict the severity of future outages considering that large amounts of information are continuously collected in the form of lessons from past events. Multiple linear regression and deep neural networks were considered for developing and testing models on the MHIDAS database. We eventually identified and discussed the inherent limitations of the techniques.

7.3. *Contribution III: Supporting the development of power grid-dedicated risk-focused solutions usable in practice.*

- This contribution aligns with objective 2.1, including sub-objectives 2.1.1, 2.1.2 and 2.1.3.
- The following articles contributed to address sub-objective 2.1.1: III, IV, VI, VII, IX.
- The following articles contributed to address sub-objective 2.1.2: I, II, VI, VII, IX, X.
- The following articles contributed to address sub-objective 2.1.3: II, VII, VIII, IX.

The reported challenges regarding continuous and reliable data acquisition, as well as the lack of methods enabling to dynamically manage the collected datasets have both hindered the application of risk-based approaches in the field of power grid management. The experience acquired during the doctoral project in those two fields enabled to suggest several ways to support the development of new risk-focused tools for better power grid management (sub-objective 2.1.1).

Several high-level challenges impacting a developed solution's performance need to be acknowledged when considering constructing new tools. When considering the use of new tools, power grid operators do indeed ideally wish to access solutions:

- better answering their needs with more precise results,
- with results provided in a timelier way,
- with results provided in a user-friendly way.

At the same time, power grid operators are also:

- limited in terms of available budget,
- particularly averse to uncertainties,
- requesting to limit the number of indicators in a solution to avoid overwhelming operators during emergency situations (i.e., human factor impacts).

We considered those constraints and addressed the problem from different perspectives.

Article III addresses the development of solutions by mainly focusing on the need to optimize the use of the resources.

It starts by suggesting using maps and large-scale remote sensing technologies such as satellite imagery to obtain a global picture of the infrastructure under review (top-down process). From there, it detects Regions of Interest (RoI) for which more detailed information is assumed to be required. A detailed inspection can then be executed on-site (bottom-up process) to confirm or reject conclusions made at the end of the top-down process. The acquired data and the final conclusions are sent back to the database via a feedback loop. After the first cycle, an implementation of a continuous data quality assessment (e.g., using the metadata-based approach developed in Article VII) will enable to assess if sufficient data is available or if further data needs to be acquired (cf. decision-making process described in Article IV).

The combination of the top-down and bottom-up processes is based on technological possibilities. It illustrates how data processing may be organized in the analytical pipeline to obtain both a global overview of the infrastructure and more local insights where required. This enables thus to improve mission prioritization and increase responsiveness while at the same time avoiding the use of resources on unnecessary inspections.

Article IV addresses the development of solutions by particularly focusing on the end-users' expectations and highlights the implications of a higher access to data with this regard.

The starting point on which Article IV is built consists of stating that the user needs that were addressed in the past when developing an analytical solution was defined by the limitations coming with little data. However, the current digitalization of society offers numerous new ways of acquiring data for an infrastructure under review. This calls thus for a restructuring of the way solutions are developed so that more end users' needs are met.

Therefore, we suggest starting by reversing the pipeline construction process and suggest thus the following steps:

- list the various possibilities available to an operator needing to make a decision,
- identify the type of information required for each possible decision,
- shape the solution in a way that enables to report that information.

Although the concept seems relatively simple, its implementation has been hindered by former pre-conceptions of solutions and by the lack of possibilities enabling to adequately retrieve the right information. The *Three-Phases Method* developed in Article VII may here again be applied to support the process, enabling thus to increase the end-users' satisfaction.

Article IX addresses the development of solutions by particularly focusing on the technical aspects of the pipeline construction, showing step-by-step which analytical action needs to be executed.

The structuring of this pipeline is based on the philosophy of both Article III and IV. It considers the user needs and the different technological possibilities for data acquisition. It also capitalizes on the result of Article VI, enabling to detect vegetation along the line. In addition, it also integrates the possibility of combining information relative to different hazards (here vegetation and wind). The analytical pipeline is then integrated into a platform, which aims to evolve and integrate information relative to additional hazards in the future. This platform reports the connections and the information flows between (1) the databases, (2) the processing unit, and (3) the different possibilities to report the information (Dashboard, APIs, etc.) on demand. In its current status, the platform enables to display interactive maps reporting the vegetation status along the power lines as well as likelihood of outage functions relative to wind induced outages.

The developments of the proposed solutions do however also imply additional analytical constraints that can strongly impact the pertinence of the provided results (sub-objective 2.1.2). In particular, they show that one needs to:

- understand how datasets can be compared when addressing one specific hazard,
- understand which datasets can be combined when addressing one specific hazard,
- understand how to combine insights regarding different hazards simultaneously considered within the scope of a risk analysis,

The part of the work summarized in Contribution II suggests ways to compare data and to deal with conflicting information when datasets are considered individually (Article VII and Article X). However, one may also need to combine insights from multiple non-synchronised datasets to extract a specific information. We faced for instance such a situation during the preparation of the work reported in Article VI. In that particular case, LiDAR point clouds captured in September 2019 were used as ground truth for detecting vegetation on satellite images captured in both September 2017 and May 2018. We dealt with those temporal discrepancies by filtering out young trees in the LiDAR point cloud and making additional assumptions relative to the vegetation growth rate over two years. The topic of dataset combination to study specific scenarios was also addressed within Articles I and II, but only on a high-level. Article IX provides, on the other hand, more detailed indications of how 2 different hazards can be analysed in parallel so that a risk analysis can eventually be executed by considering them simultaneously.

The development of a solution to be used by grid operators also requires the acknowledgment of real-world constraints directly impacting the likelihood of having the proposed solutions indeed used in by utilities in operation (sub-objective 2.1.3).

The application of the reinforcement action suggested in the metadata-based approach (Article VII) needs first to be applied case by case. The budget available for a risk analysis project might vary considerably from organization to organization. Smaller companies are for instance more likely to face a lack of financial resources and knowledge in their teams than bigger utilities. This directly impacts the reachable level of insights and shall thus be clarified to make sure that the stakeholders' expectations align with the project's context.

Real-world situations also imply that ideal conditions might not be met. For instance, heterogeneous technologies used in the development of a system might lead to technological hardware and/or software compatibility issues, as is highlighted in Article II. This problem is also raised regarding the use of standards in Article VII and VIII. Different standards may be chosen by different stakeholders, for instance in the way of reporting metadata. Lack of compliance with any standard may also be faced, therefore directly amplifying the reported problems. There is thus a need to emphasize the importance of coordinated application of standards when reporting metadata, which is also supported by the present work.

Finally, the success of the creation of a solution is also strongly dependent on aspects such as:

- the intensity of the demand in the market,
- the existence of potential competitors,
- the available analytical capacities of the involved stakeholders,
- the available software development capacities used to build the product,
- a strong business plan ensuring that the solution is financially viable,
- the capacity of regulatory authorities to acknowledge the possible technological evolutions and adapt the regulation,
- etc.

Such types of insights were acknowledged but not largely detailed in the development of Article IX. However, they have been taken into consideration in the application of the GridEyeS feasibility study (successfully concluded) and in the application of the GridEyeS demonstration project (application under review).

7.4. Contribution IV: Supporting value generation out of historical data and supporting decision-making regarding future data captures.

- This contribution is in line with objective 2.2.
- The following articles contributed to address objective 2.2: I, II, III, IV, V, VI, VII, VIII, IX, X

All the articles prepared for the present thesis provided recommendations usable by power grid operators to better generate value out of the data they have and/or better support decision-making regarding future data capture.

In Article I, we suggested a data combination that could help operators to predict cracks in power poles. In Article II, we paved the way for multiple research avenues by suggesting combination of heterogeneous datasets in additional specific use-cases and indicated where data relevant for a Norwegian environment could be accessed. The solutions developed and presented via Articles III, IV, and IX provide recommendations on how and when the different datasets could be integrated in the analytical pipelines to optimize the management of the grids. The outcome of the analytical pipelines enables in particular to objectively decide how and where future data acquisition should be planned and prioritized based on available resources, existing data and previous risk analyses executed in specific areas. These solutions thus enable data-driven and risk-based decision-making instead of calendar-based, enabling to increase the grid's resilience.

Article V gave suggestions on how outage reports may be used to predict the severity of future outages based on lessons learned with previous major accidents. Article VI indicated how LiDAR data could be used to train models detecting vegetation along the lines using additional satellite data, therefore being able to better predict the probability of tree-fall-related outages. The article also suggests a way to display the results, which is later integrated into the development of the platform proposed in Article IX.

The approach suggested in Article VII and applied in Article VIII gives the opportunity to grid operators to assess the quality of, compare and rank any dataset they consider using for risk analysis. This assessment can be executed by indicating how each dataset can inform the parameters relevant in the risk analysis. Depending on the chosen decision rules, one may also consider using the veracity indicator suggested in Article X, which enables to control the degree of agreement among the sources when different sources are considered simultaneously.

Article VIII also reports all the types of data sources that have been identified as relevant to estimating the probability of a tree fall. Similarly to Article II, Article VIII emphasizes the need to facilitate access to data and provides complementary indications on where public data could be found. It also reports a list of tools that could be used to work with the different datasets, which is further reported in Table A of Appendix A.

7.5. Contribution V: Enabling vegetation management improvement.

- This contribution is in line with objective 2.3.
- The following articles contributed to address objective 2.3: II, III, IV, VI, VII, VIII, IX, X.

Vegetation-management represented the main case-study used to illustrate the applicability and pertinence of the contributions made in the present thesis. A bow-tie analysis was used to detail the scenario of a tree falling on a power line in Articles II and IV. Computer-vision was used to automatically detect vegetation along the power grid using satellite images in Article VI. The applicability of the metadata-based approach proposed in Article VII was first assessed for vegetation characterization using simulated data in that same work, and then assessed on an identical case-study using real-world data in Article VIII. As such, the different general recommendations and comments made in the previous contributions also apply when focusing on the specific case of vegetation management. The main implications for vegetation management along the lines is thus only shortly summarized in the following, also acknowledging that it partially overlaps with the previous sections.

The different solutions and methods presented in this work will enable grid operators to:

- diversify the data sources they used to consider to assess the status of vegetation along the lines, therefore increasing their resilience to loss of information and reducing their dependency on specific data providers,
- increase their understanding of the mechanisms behind vegetation-related outages, and therefore,
- affine their understanding of scenarios considered in the risk analysis by adequately identifying and characterizing relevant parameters,
- increase the frequency of risk estimations by increasing the use of remote-sensing technologies,
- make case-specific risk estimations/improvements by gathering local data, avoiding use of averaged values and reducing thus uncertainty around risk estimation,
- discover previously unconsidered risky areas,

The final results provided in Article VI and enabling to better predict the probability of outage due to a tree fall on a power line is based a quantitative indicator called the *Tree Density Index*. This indicator considers both the proximity and the density of the vegetation along the power line, therefore optimally summarizing the vegetation status along the infrastructures. Combining such insights with the possibility of increasing the number of data sources and datasets will thus facilitate vegetation management mission prioritization and increase the responsiveness of the utilities when dealing with this hazard.

8. Discussion

8.1.Discussion of the 1st Contribution: Implications of new data acquisition opportunities in QRAs.

The research activities of the present doctoral project have been applied in the field of power grid management. The most detailed analyses executed during the Ph.D. have focused on the case of vegetation along power lines in a Norwegian environment. The type of data sources reported for that work would very likely be valuable when focusing on other environments and/or on different hazards. However, alternative data sources may also be relevant for such new tasks and a new benchmark of available data sources would therefore be required to complete the existing list. Such a recommendation is even more meaningful when considering the execution of risk analysis applied on other types of infrastructures. This comment is also applicable regarding the links provided for publicly available datasets, as they may be less relevant for analyses applied in a non-Norwegian environment.

One may also discuss the pertinence of considering the variety of data sources that have been suggested. Indeed, the limited value that may sometimes be provided by the processing of new types of datasets may not compensate for the processing costs. Therefore, investigating the potential value of a new dataset needs to be based on the experience of the dataset analysts and the power grid operators involved in the project. More experience at the beginning of the doctoral project could for instance have helped to understand that looking for the prediction of cracks in power poles (Article I) may be a secondary topic considering the low numbers of disturbances related to pole cracks.

Multiple reasons can lead to similarly inadequate choices in a more industrial context, such as:

- the lack of human resources,
- the lack of financial resources,
- the lack of technological resources,
- the lack of familiarization with the types of data sources (gap in analytical knowledge).

Understanding where the bottleneck is before starting a risk analysis project enables thus to optimize the decision-making process. This could, for instance, enable to decide that investing in operator competency development may be more strategical than spending resources on acquiring additional datasets. In fact, acquiring data with a higher resolution may for instance not always be meaningful. The definition of the ideal resolution is indeed defined during the parameter characterization, (2nd reinforcement action of the metadata-based approach (Article VII)) and going beyond that resolution would simply result in a waste of resources.

8.2. *Discussion of the 2nd Contribution: Benefits and limitations of conventional risk assessment frameworks augmentation.*

The definition of the level of convergence suggested as a first reinforcement action in the metadata-based approach eventually aims at maximizing the possibilities to reach the initial objectives by using the best level of insights possible for the risk analysis considering the available resources. There may, however, not be a single level of convergence when managing a discretized large-scale infrastructure. Further work would then be required to assess how the choice of multiple levels of convergence relative to multiple items of interest should simultaneously be handled in the management of such an infrastructure. The characterization suggested in the second reinforcement action only makes sense if one knows how to generate meaningful models based on the reported parameters. This is a labour-intensive task requiring a diversity of competencies, which therefore strongly calls for implementing cross-disciplinary teams. However, we believe that the suggested processing remains the best way of defining reference values to be used for the later quality assessment of datasets considered for a risk analysis.

The metadata-based method then uses this information to dynamically rank any dataset that originates from data sources identified during the second reinforcement action. The application of this method enables to provide more certain results for a risk analysis, potentially at a lower cost. However, one should also be aware of the high efforts that may be required during the implementation phase of the method. Furthermore, additional work in field-specific conditions is required to ensure that case-specific constraints (definition of weights in the method, adaptation to different standard choices, etc.) are adequately addressed for the analysis of the infrastructures under review. This includes for instance the definition of decision rules (e.g., use of veracity indicator) to determine if and how multiple datasets should be considered simultaneously within a risk analysis.

Both the metadata-based approach and the use of the veracity indicators enable to characterize and/or reduce the uncertainty relative to data integration. By doing so, they allow reducing the global level of uncertainty in the risk analysis. However, other sources of uncertainties originating from the choice of the models selected for the quantitative risk analysis still need to be addressed to increase the pertinence of the provided results.

The metadata-based approach, the computer-vision-based solutions exploited in this work and the work focusing on outage severity prediction illustrate in various ways how the machine-learning field can support more efficient risk analysis. However, one should remember that the performances of machine learning-based solutions are directly impacted by both the nature and quality of the data used to train the models and the way the models are developed. The different studies reported in the present thesis highlight thus here again the need of good practices in data management processes, as well as the need to diversify the panel of experts involved in the solution development. Furthermore, the presented applications do obviously only report a limited panel of ways machine learning techniques can be used to support better risk analysis. Numerous alternative approaches may additionally be explored to make new risk-focused solutions more accurate, more reactive and more dynamic.

8.3. Discussion of the 3rd Contribution: Benefits and limitations of solution development-related recommendations

The insights obtained by addressing the objectives in block 1 have enabled the development of different risk-focused solutions usable in the field of power grid management. Although addressing the objective of solution development from different perspectives, the different studies presented in this thesis do logically globally converge to the same direction. Indeed, the developed solution:

- Are primarily centered around user-needs,
- Acknowledge technical and technological constraints,
- Acknowledge the existence of the main practical constraints faced by each stakeholder involved in the solution development.

This thesis thus contributes to developing new tools enabling better risk analysis in power grids.

One may be reminded that the proposed solutions assume an extended data access. The presented results showed that such an assumption can be valid when addressing the problem of vegetation along power lines. However, data may not be extensively available when focusing on other hazards impacting the grid, therefore potentially limiting the extension possibilities of the solutions we proposed.

Regular benchmarks of technologies and technics are thus continuously required to maximize the solutions' applicability. Executing such benchmarks might in addition be useful to explore new dataset combination possibilities. However, combination possibilities also require further research enabling to understand which type of risk-related insights can be obtained. Knowing which dataset combination may be valuable is a task requiring both experience and creativity, calling thus once more for cross-disciplinary teams. It also requires to further investigate the field of data fusion to implement even better information control processes.

Further exploring the field of data fusion will enable to assess how additional decisions rules may be integrated in the proposed solutions. This challenging task requires clarity and flexibility to enable future adaptations potentially required due to context modifications.

The provided solutions also show multiple additional evolution opportunities. Indeed, the present thesis mainly focused on vegetation; but numerous additional hazards must be addressed to provide more comprehensive solutions. Additionally, and although the management of substations has been explored during the Ph.D., the main type of elements of the grid that was analysed is overhead power lines. Further elements such as power production sites and power consumption nodes would also have to be integrated to increase the value of the developed solutions. Finally, the work presented so far mainly focused on optimizing data integration for better risk analysis, reducing that way the level of uncertainty when executing the risk analysis. However, the risk quantification was yet left aside. Such a task is obviously essential and will be part of future works.

The proposed solutions were developed in an industrial context, considering real-world constraints impacting the entire information pipeline (i.e., data capture, data transmission, data pre-processing, information processing, results transmission). They were also developed in close collaboration with the end-users and conceptually validated by them. However, they are still only part of preliminary studies. Turning them into viable commercial products consists of much larger projects requiring extensive market analysis and a complete software production plan that needs to be defined on company-level.

8.4. Discussion of the 4th Contribution: Extension opportunities for value generation based on historical data and future data capture.

The work presented in this thesis provides multiple recommendations to grid operators on how to generate value out of the historical data they can access (internal database, open-access services, commercial data providers, etc.). It does so by reporting multiple use-cases in which the reported datasets can be used, also reporting how the data may be processed.

The metadata-based approach especially enables utilities to simultaneously assess the value from a risk analysis perspective of all the data they may access at a given point of time. On this basis, grid operators can thus easily determine the parts of the network for which a new data acquisition is the most urgent. In addition, the power grid operators have the possibility to select the most meaningful data acquisition method by considering the informative potential of each type of dataset together with their associated acquisition and processing costs.

However, the parametrization required for the method's application remains a complex task. Utilities would still require specific expertise relative to each of the relevant hazards (e.g., forest experts for vegetation, meteorologists for weather) to implement and continuously control the results of such a solution. This may imply hiring external consultants also coming at some costs and complicating the solutions' implementation.

Finally, the solutions that have been provided with the present doctoral work have not yet been implemented in software solutions usable by operators in operation. This remains a stumbling block that needs to be removed to maximize the support to utilities provided by this thesis.

8.5.Discussion of the 5th Contribution: Effective optimization level of vegetation management.

The different studies reported in this Ph.D. have enabled to obtain a comprehensive overview of the mechanisms involved in the problem represented by vegetation existing along power lines. They have also enabled to report the different options available to the operators when having to deal with that hazard.

Both the capacity to exploit new types of technologies (e.g., satellite images) and the possibility to objectively compare the informative potential of a dataset with its acquisition/processing costs provide thus new opportunities to optimize vegetation management along power lines.

However, the different studies have also shown that further work is required to increase the performances and applicability of the methods proposed in the different articles. In addition, vegetation-related quantitative risk modelling still needs to be developed in order to optimally support utilities. Finally, the number of industrial-level software applications nowadays able to handle vegetation analysis in a cost-efficient and time-efficient way remains particularly limited. More work is thus also required in that direction. The existence of this gap has motivated the preparation of the GridEyeS demonstration project application, currently under review by the European Space Agency.

9. Conclusions & Further Work

9.1. Conclusion

The digitalization of our society supports the development of tools enabling to make more dynamic and better risk-informed decisions. This is also true in the field of power grid management, where the increase of data acquisition possibilities represents a strong opportunity to reach higher operational performances. Indeed, new technologies and new data processing technics support power grid operators by facilitating access to more, better and more frequent infrastructure status data. This contrasts with the lack of data and the lack of adapted methods that have previously hindered the development of efficient risk-based solutions.

However, the digitalization is also synonym of numerous challenges. For example, information may originate from a multitude of heterogeneous data sources, therefore implying heterogeneity in the informative potential of the datasets, as well as different processing constraints and costs. Such challenges may, in turn, be problematic for the application of risk-based decision-making. In fact, there is nowadays a lack of solutions enabling to deal with heterogeneous datasets in the field of risk analysis in general, de facto limiting the number of solutions in the particular case of power grid management.

This Ph.D. had two main ambitions based on the acknowledgment of the presented situation:

(1) Supporting the development and implementation of Dynamic Risk Analysis in practice.
(2) Supporting power grid operations in real-world circumstances by enabling better risk-informed decisions.

Those objectives were addressed by focusing on optimizing the data flow in the information pipelines. To do so, it first took advantage of the Industrial Ph.D. scheme configuration to get in close contact with multiple utilities and power grid operators. This enabled to adequately identify their expectations, their working process and benefit from their field expertise. It especially enabled to understand how data may commonly be captured during their daily operations. The research activities that were then executed within multiple real-world industrial projects then enabled to acquire knowledge and experience with various technologies and technics for data capture and processing.

Combining this know-how with an academic background in the field of risk analysis and a systems engineering mindset enabled this thesis to provide the following contributions:

(1) **Alternative data sources.**
We benchmarked and reported alternative data sources that could be considered for integration to increase data access within quantitative risk analyses. This also consisted of referencing multiple open access data sources & tools to favour reproducibility and knowledge-sharing.

(2) **Augmentation of conventional risk assessment frameworks.**

We first made several recommendations to reinforce conventional risk assessment frameworks. Those mainly aimed at ensuring that all the elements required for an efficient risk analysis are adequately identified. We then proposed several ways to compare and/or combine multiple heterogeneous datasets to reduce uncertainties due to data integration. We also investigated multiple ways of using the machine learning field to support more automatic and dynamic risk analyses, both in theory and in practice.

(3) **Development of power grid-dedicated risk-focused solutions.**

We provided multiple user-driven propositions of solution development to be used for the construction of risk-focused tools usable by power grid operators in future operations.

(4) **Data use and acquisition optimization in power grid management.**

We provided multiple recommendations enabling power grid operators to better exploit the historical data they can already access. We also provided multiple recommendations enabling them to plan future data acquisition more efficiently.

(5) **Vegetation management improvement.**

We facilitated access to alternative data sources, enabling diversification of information origin in vegetation management. This makes the vegetation-related risk analysis more complete and more resilient to data source loss. Additionally, we showed how the different datasets may be integrated and combined to favor better vegetation-related risk depiction in the context of power grid management.

Eventually, this thesis enabled to contribute to both the development and practical implementation of dynamic risk analysis in real-world configurations by diversifying data acquisition opportunities and reducing the inertia coming with data processing. It also enabled to support the development and implementation of efficient risk-focused tools usable in the context of power grid management by highlighting how data could be better combined, compared and integrated into various information processing pipelines, especially focusing on vegetation management.

9.2. Further Work

The different limitations that have been mentioned in the discussion sections of both the articles presented in Part II and in the previous sections suggest various opportunities for further work. We center our suggestions around the five main contributions provided in this thesis.

(1) **Alternative data sources.**

There will be a need to look for alternative data capturing processes as long as spatiotemporal resolution discrepancies will be observed between the reference values reported during the parameter characterization and the properties of the datasets used to analyse a scene. Regular benchmarks of new technologies, tools and data processing technics are thus continuously required in the future to ensure that the best data capture modes are known to the various stakeholders involved in the development of software solutions. A process that may furthermore be applied while waiting for that gap to be filled is -in line with our contributions- to efficiently combine proactive data acquisition procedures (leading data capture – e.g., automatic satellite image analysis) with reactive data acquisition procedures (lagging data capture – e.g., on-demand data capture in the field). "Efficiently" implies continuously and dynamically finding the right trade-off between analytical performances (reactivity, accuracy, precision, etc.) and available resources (time, money, people, material, etc.). This corresponds to an optimization problem that needs to solved case by case, depending on organization and infrastructure.

(2) **Augmentation of conventional risk assessment frameworks.**

The approach suggested for augmenting conventional risk assessment frameworks is particularly dependent on the quality of data reporting. Therefore, it called for a better compliance with standards when generating data. Our work could thus be used to illustrate to the stakeholders involved in data registration how following guidelines is relevant for that task. Further work should however focus on making the proposed approach more resilient to the consideration of datasets not conforming with existing standards and protocols. In line with that suggestion, further work may focus on making the suggested method less dependent to the assumptions made during the approach development.

We also suggested using the metadata-based approach to take decisions regarding future data acquisition. However, this suggestion is only based on the default maximum potential of knowledge provided by the considered data sources. Further work should also look into the data acquisition costs and frequencies relative to each data source to optimally select the data sources to consider in the implementation of future inspection procedures.

Further work should obviously also focus on the risk calculations by effectively developing the quantitative risk models that will enable to optimally exploit the data that has been made available. With this in mind, future work will especially have to understand the implications of the interactions between different grid elements on which a risk analysis will have been applied. Such an investigation is particularly relevant in the case of cascading outages.

Finally, we also suggest investigating how machine learning methods may further support the development and implementation of dynamic risk analysis. Such investigations may for instance be particularly helpful to capture the interactions between multiple hazards, enabling therefore to better link causes and consequences in the risk analyses.

(3) **Development of power grid-dedicated risk-focused solutions.**

The solutions we have suggested were mainly focusing on the topic of vegetation management. As we know, multiple additional hazards also represent a risk to power grids. Further work may thus investigate how the proposed solutions may be extended to other hazards, looking in particular to the interactions existing between those. As we also pointed out in the thesis, the provided results are part of preliminary studies and the solutions have not yet been coded to be usable in a real software. Further work is thus planned to use the suggested solutions to develop commercial products.

Additional non-analytical tasks also need to be addressed to ensure the successful implementation of the proposed solutions. Market analysis, competitor landscapes investigation and business case reinforcements are for instance required to ensure that the developed solution will financially be viable in the future. A pedagogical work is furthermore required to adequately communicate the results to regulatory organizations to enable the adaptation of existing regulations, which may sometimes hinder the implementation of innovative processes.

(4) **Data use and acquisition optimization in power grid management.**

We obviously also suggest to further closely work with utilities to ensure that their needs are always adequately addressed. This implies further investigating how to get additional value out of the data they can access. For instance, further work may look into new dataset combination possibilities to identify the parameters playing a role in the risk calculations.

(5) **Vegetation management improvement.**

Further work may simply extend the application of the metadata-based approach that focused on vegetation management. This first implies further characterizing the parameters and the informative potential of the reported data sources for the two other dimensions influencing the *probability* of a tree fall (intrinsic tree stability and external forces). This also implies investigating the *consequence* dimension for such a *critical event*. Finally, this also implies investigating the second important *scenario*, focusing on the problem of vegetation growth under the power lines.

The industrial context in which this thesis has been executed has favoured the outcome of the provided contributions. It is therefore a strong reminder of the importance of braking barriers - between industry and academia, as well as between heterogeneous fields of expertise – to successfully close a project.

References

AEA technology - Major hazards assessment unit, 2003. MHIDAS – Major Hazard Incident Data Service. UK.

Andrews, D.F., 1974. A Robust Method for Multiple Linear Regression. Technometrics 16, 523–531. https://doi.org/10.1080/00401706.1974.10489233

Aven, T., 2012. The risk concept—historical and recent development trends. Reliab. Eng. Syst. Saf. 99, 33–44. https://doi.org/http://dx.doi.org/10.1016/j.ress.2011.11.006

Aven, T., 2008. Risk Analysis: Assessing Uncertainties beyond Expected Values and Probabilities. John Wiley & Sons Ltd, West Sussex, England.

Aven, T., Krohn, B.S., 2014. A new perspective on how to understand, assess and manage risk and the unforeseen. Reliab. Eng. Syst. Saf. 121, 1–10.

Bell, K., Chen, Y., Chowdhury, B., Dobson, I., Hines, P., Papic, M., Miller, S., Zhang, P., 2012. Risk Assessment of Cascading Outages : Methodologies and Challenges. IEEE Trans. Power Syst. 27, 631–641.

Brown, R.E., Spare, J.H., 2004. Asset Management, Risk, and Distribution System Planning, in: IEEE PES Power Systems Conference and Exposition, 2004. IEEE, p. 6. https://doi.org/10.1109/PSCE.2004.1397445

Brynjolfsson, E., Mitchell, T., 2017. What can machine learning do? Workforce implications. Science (80-.). 358, 1530–1534. https://doi.org/10.1126/science.aap8062

Bucelli, M., Paltrinieri, N., Landucci, G., 2018. Integrated risk assessment for oil and gas installations in sensitive areas. Ocean Eng. 150, 377–390. https://doi.org/10.1016/j.oceaneng.2017.12.035

Buetti-Dinh, A., Galli, V., Bellenberg, S., Ilie, O., Herold, M., Christel, S., Boretska, M., Pivkin, I. V., Wilmes, P., Sand, W., Vera, M., Dopson, M., 2019. Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. Biotechnol. Reports 22, e00321. https://doi.org/10.1016/j.btre.2019.e00321

Canadian Standards Association (CSA), 1997. Risk Management: Guideline for Decision Makers. CAN/CSA-Q850-97 (R2009). Ottawa, Canada.

Catrinu, M.D., Nordgard, D.E., 2011. Integrating risk analysis and multi-criteria decision support under uncertainty in electricity distribution system asset management. Reliab. Eng. Syst. Saf. 96, 663–670. https://doi.org/10.1016/j.ress.2010.12.028

Ciapessoni, E., Cirio, D., Member, S., Kjølle, G., Member, S., Massucco, S., Member, S., Pitto, A., Sforna, M., 2016. Probabilistic Risk-Based Security Assessment of Power Systems Considering Incumbent Threats and Uncertainties. IEEE Trans. Smart Grid 7, 2890–2903. https://doi.org/10.1109/TSG.2016.2519239

Creedy, G.D., 2011. Quantitative risk assessment: How realistic are those frequency assumptions? J. Loss Prev. Process Ind. 24, 203–207. https://doi.org/10.1016/j.jlp.2010.08.013

Delvosalle, C., Fievez, C., Pipart, A., Debray, B., 2006. ARAMIS project: A comprehensive methodology for the identification of reference accident scenarios in process industries. J. Hazard. Mater. 130, 200–219. https://doi.org/10.1016/j.jhazmat.2005.07.005

Doostan, M., Sohrabi, R., Chowdhury, B., Francisco, S., 2019. A Data-Driven Approach for Predicting Vegetation-Related Outages in Power Distribution Systems 1–13.

Eggum, E., 2019. Rapport Nr. 29-2019 - Avbrotsstatistikk 2018. Oslo.

Enedis, 2020. Enedis - En Bref [WWW Document]. URL https://www.enedis.fr/profil

Enedis, 2017. Chiffres Clés.

Eurelectric, 2020. Distribution Grids in Europe - Facts and Figures 2020.

European Commission, 2017a. HyRiM - Hybrid Risk Management for Utility Networks [WWW Document]. Community Res. Dev. Inf. Serv. URL https://cordis.europa.eu/project/id/608090

European Commission, 2017b. GARPUR - Generally Accepted Reliability Principle with Uncertainty modelling and through probabilistic Risk assessment [WWW Document]. Community Res. Dev. Inf. Serv. URL https://cordis.europa.eu/project/id/608540

European Commission, 2016. iTesla - Innovative Tools for Electrical System Security within Large Areas [WWW Document]. Community Res. Dev. Inf. Serv. URL https://cordis.europa.eu/about/en

European Commission, 2015. UMBRELLA - Toolbox for Common Forecasting, Risk assessment, and Operational Optimisation in Grid Security Cooperations of Transmission System Operators (TSOs) [WWW Document]. Community Res. Dev. Inf. Serv. URL https://cordis.europa.eu/project/id/282775/de

European Commission, 2014a. AFTER - A Framework for electrical power sysTems vulnerability identification, dEfense and Restoration [WWW Document]. Community Res. Dev. Inf. Serv. URL https://cordis.europa.eu/project/id/261788

European Commission, 2014b. COMMISSION REGULATION (EU) No 651/2014.

European Commission, 2014c. Communication from the Commission - Framework for State aid for research and development and innovation (2014/C 198/01). Off. J. Eur. Union 1–29.

European Network of Transmission System Operators for Electricity (ENTSO-E), 2019. STATISTICAL FACTSHEET 2018.

European Network of Transmission System Operators for Electricity (ENTSO-E), 2010. Operational Handbook – Policy 5 : Emergency Operations.

FERC and NERC, 2011. Report on Outages and Curtailments During the Southwest Cold Weather Event of February 1-5, 2011.

Froomkin, A.M., Kerr, I., Pineau, J., 2019. WHEN AIs OUTPERFORM DOCTORS: CONFRONTING THE CHALLENGES OF A TORT-INDUCED OVER-RELIANCE ON MACHINE LEARNING. Ariz. Law Rev. 61, 33–100.

GARPUR Project, 2017. GARPUR - Deliverables [WWW Document]. URL https://www.sintef.no/projectweb/garpur/deliverables/

Gazzea, M., Pacevicius, M., Dammann, D.O., Sapronova, A., Lunde, T.M., Arghandeh, R.,

2021. Automated Power Lines Vegetation Monitoring using High-Resolution Satellite Imagery. Trans. Power Deliv. 1–9.

Giraud-carrier, C., 2008. Metalearning - A Tutorial.

Goodfellow, I.J., Bengio, Y., Courville, A., 2016. Deep learning, The MIT Press. Citeseer, Cambridge, Massachusetts, US.

Guikema, S.D., Davidson, R.A., Liu, H., 2006. Statistical Models of the Effects of Tree Trimming on Power System Outages 21, 1549–1557.

Haes Alhelou, H., Hamedani-Golshan, M., Njenda, T., Siano, P., 2019. A Survey on Power System Blackout and Cascading Events: Research Motivations and Challenges. Energies 12, 682. https://doi.org/10.3390/en12040682

Hansen, H., 2018. Rapport Nr. 64-2018 - Avbrotsstatistikk 2017. Oslo.

Haskins, C., 2008. Systems engineering analyzed, synthesized, and applied to sustainable industrial park development. Syst. Eng.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Second Edi. ed. Springer, New York. https://doi.org/10.1007/b94608_4

INCOSE, 2021a. About Systems Engineering [WWW Document]. URL https://www.incose.org/about-systems-engineering

INCOSE, 2021b. Systems Engineering [WWW Document]. URL https://www.incose.org/systems-engineering

IRGC - International Risk Governance Council, 2009. Risk Governance Deficits. An analysis and illustration of the most common deficits in risk governance. Geneva.

Isaias Task Force, 2020. TROPICAL STORM ISAIAS 30-DAY REPORT.

ISO - International standardization organization, 2018. ISO 31000:2018 - Risk management - Guidelines. Geneva, Switzerland.

Johansson, E., Uhlen, K., Kjølle, G., Toftevaag, T., 2011. Reliability evaluation of wide area monitoring applications and extreme contingencies. 17th Power Syst. Comput. Conf. PSCC 2011.

Kalantarnia, M., Khan, F., Hawboldt, K., 2009. Dynamic risk assessment using failure assessment and Bayesian theory. J. Loss Prev. Process Ind. 22, 600–606. https://doi.org/10.1016/j.jlp.2009.04.006

Kaplan, S., Garrick, B.J., 1981. On The Quantitative Definition of Risk. Risk Anal. 1, 11–27. https://doi.org/10.1111/j.1539-6924.1981.tb01350.x

Khakzad, N., Khan, F., Amyotte, P., 2012. Dynamic risk analysis using bow-tie approach. Reliab. Eng. Syst. Saf. 104, 36–44. https://doi.org/10.1016/j.ress.2012.04.003

Khuntia, S.R., Rueda, J.L., Meijden, M.A.M.M. Van Der, 2017. Smart Asset Management for Electric Utilities: Big Data and Future, in: WCEAM 2017.

Lee, S., Landucci, G., Reniers, G., Paltrinieri, N., 2019. Validation of dynamic risk analysis supporting integrated operations across systems. Sustain. 11, 1–25. https://doi.org/10.3390/su11236745

Libertson, F., 2021. Competing socio-technical narratives in times of grid capacity challenges: the representative case of Sweden. Energy. Sustain. Soc. 11, 1–13. https://doi.org/10.1186/s13705-021-00279-4

Masson-Delmotte, V., Pörtner, H.-O., Skea, J., Zhai, P., Roberts, D., Shukla, P.R., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J.B.R., Chen, Y., Zhou, X., Gomis, M.I., Lonnoy, E., Maycock, T., Tignor, M., Waterfield, T., 2018. Global Warming of 1.5°C: Summary for Policymakers, Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change,. Geneva, Switzerland,. https://doi.org/10.1017/CBO9781107415324

Matikainen, L., Lehtomäki, M., Ahokas, E., Hyyppä, J., Karjalainen, M., Jaakkola, A., Kukko, A., Heinonen, T., 2016. Remote sensing methods for power line corridor surveys. ISPRS J. Photogramm. Remote Sens. 119, 10–31. https://doi.org/10.1016/J.ISPRSJPRS.2016.04.011

Meadows, D.H., 2009. Thinking in Systems - A Primer, Earthscan. London.

Mittal, S., Vaishay, S., 2019. A survey of techniques for optimizing deep learning on GPUs. J. Syst. Archit. 99, 101635. https://doi.org/10.1016/j.sysarc.2019.101635

Motter, A.E., Myers, S.A., Anghel, M., Nishikawa, T., 2013. Spontaneous synchrony in power-grid networks. Nat. Phys. 9, 191–197. https://doi.org/10.1038/nphys2535

Narasimhan, S., Jordache, C., 2000. Data Reconciliation and Gross Error Detection. An Intelligent Use of Process Data. Gulf Publishing Company, Houston, Texas.

NERC, 2018. Hurricane Irma Event Analysis Report 1–33.

Nguyen, V.N., 2019. Advancing Deep Learning for Automatic Autonomous Vision-based Power Line Inspection. UiT - The Arctic University of Norway.

Nguyen, N. Van, Jenssen, R., Roverso, D., 2018. Automatic Autonomous Vision-based Power Line Inspection : A Review of Current Status and the Potential Role of Deep Learning. Int. J. Electr. Power Energy Syst.

Nordgård, D.E., 2010. Risk Analysis for Decision Support in Electricity Distribution System Asset Management. Norwegian University of Science and Technology (NTNU).

NORSOK, 2010. Standard Z-013, Risk and Emergency Preparedness Assessment. Lysaker, Norway.

OECD, 2015. Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measur. ed. OECD Publishing, Paris. https://doi.org/http://dx.doi.org/10.1787/9789264239012-en

Pacevicius, M., A. Ramos, M., Paltrinieri, N., 2020. Optimizing Technology-based Decision-support for management of Infrastructures under risk: The Case of Power Grids, in: Proceedings of the 30th ESREL-15th PSAM. Venice, p. 8.

Pacevicius, M.F., Ramos, M., Roverso, D., Eriksen, C.T., Paltrinieri, N., 2022. Managing Heterogeneous Datasets for Dynamic Risk Analysis of Large-Scale Infrastructures. Energies 15, 40. https://doi.org/10.3390/en15093161

Paltrinieri, N., Dechy, N., Salzano, E., Wardman, M., Cozzani, V., 2013a. Towards a new approach for the identification of atypical accident scenarios. J. Risk Res. 16, 337–354. https://doi.org/10.1080/13669877.2012.729518

Paltrinieri, N., Dechy, N., Salzano, E., Wardman, M., Cozzani, V., 2012. Lessons Learned from Toulouse and Buncefield Disasters: From Risk Analysis Failures to the Identification of Atypical Scenarios Through a Better Knowledge Management. Risk Anal. 32. https://doi.org/10.1111/j.1539-6924.2011.01749.x

Paltrinieri, N., Khan, F., 2016. Dynamic Risk Analysis in the Chemical and Petroleum Industry: Evolution and Interaction with Parallel Disciplines in the Perspective of Industrial Application. Elsevier.

Paltrinieri, N., Khan, F., Amyotte, P., Cozzani, V., 2013b. Dynamic approach to risk management: Application to the Hoeganaes metal dust accidents. Process Saf. Environ. Prot. 92, 669–679. https://doi.org/10.1016/j.psep.2013.11.008

Paltrinieri, N., Khan, F.I., 2020. Dynamic risk analysis—Fundamentals, in: Khan, F.I., Amyotte, P.R.B.T.-M. in C.P.S. (Eds.), Advanced Methods of Risk Assessment and Management. Elsevier, pp. 35–60. https://doi.org/https://doi.org/10.1016/bs.mcps.2020.04.001

Pasman, H., Reniers, G., 2014. Past, present and future of Quantitative Risk Assessment (QRA) and the incentive it obtained from Land-Use Planning (LUP). J. Loss Prev. Process Ind. 28, 2–9. https://doi.org/10.1016/j.jlp.2013.03.004

Perkin, S., 2018. Real-time weather-dependent probabilistic reliability assessment of the icelandic power system. Reykjavík University.

Perkin, S., Svendsen, A.B., Tollefsen, T., Honve, I., Baldursdottir, I., Stefansson, H., Kristjansson, R., Jensson, P., 2017. Modelling weather dependence in online reliability assessment of power systems. Proc. Inst. Mech. Eng. Part O J. Risk Reliab. 231, 364–372. https://doi.org/10.1177/1748006X17694951

Pimentel, B.A., de Carvalho, A.C.P.L.F., 2020. A Meta-learning approach for recommending the number of clusters for clustering algorithms. Knowledge-Based Syst. 195, 105682. https://doi.org/10.1016/j.knosys.2020.105682

Radmer, D.T., Kuntz, P.A., Christie, R.D., Venkata, S.S., Fletcher, R.H., 2002. Predicting Vegetation-Related Failure Rates for Overhead Distribution Feeders 17, 1170–1175.

Raussand, M., 2011. RISK ASSESSMENT - Theory, Methods, and Applications, Handbook on the Toxicology of Metals: Fourth Edition. Wiley.

Reinsel, D., Gantz, J., Rydning, J., 2018. Data Age 2025: The Digitization of the World From Edge to Core. Seagate, IDC 28.

Research Council of Norway, 2021. Industrial PhD Scheme – Doctoral Projects in Industry [WWW Document]. URL https://www.forskningsradet.no/en/call-for-proposals/2019/industrial-ph.d.-scheme--doctoral-projects-in-industry/

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 386–408. https://doi.org/10.1037/h0042519

Samuel, A.L., 1959. Some Studies in Machine Learning Using the Game of Checkers. IBM J. Res. Dev. 3, 210–229. https://doi.org/10.1147/rd.33.0210

Sand, K., Gjerde, O., Nordgård, D.E., Catrinu, M., 2007. Current risk exposure in the distribution sector. Initial study. SINTEF Energy Res. Trondheim 6576.

Sand, K., Kjolle, G., Bilberg, J., 1989. Reliability aspects concerning distribution system expansion planning, in: 10th International Conference on Electricity Distribution, 1989: CIRED 1989. pp. 530–534.

Sforna, M., Delfanti, M., 2006. Overview of the events and causes of the 2003 Italian blackout, in: 2006 IEEE PES Power Systems Conference and Exposition, PSCE 2006 - Proceedings. pp. 301–308. https://doi.org/10.1109/PSCE.2006.296323

Smith, A.B., 2021. 2020 U.S. billion-dollar weather and climate disasters in historical context [WWW Document]. URL https://www.climate.gov/news-features/blogs/beyond-data/2020-us-billion-dollar-weather-and-climate-disasters-historical

Sullivan, B.K., Malik, N.S., 2021. 5 Million Americans Have Lost Power From Texas to North Dakota After Devastating Winter Storm [WWW Document]. Time. URL https://time.com/5939633/texas-power-outage-blackouts/#:~:text=5 Million Americans Have Lost,on Feb. 15%2C 2021.

Task Force on Power Outages - Eurelectric, 2006. Impacts of Severe Storms on Electric Grids.

U.S.-Canada Power System Outage Task Force, 2004. Final Report on the August 14, 2003 Blackoutin the United States and Canada: Causes and Recommendations. https://doi.org/10.1021/ac203394r.A

United States Department of Energy, O. of E.D. and E.R., 2013. Comparing the Impacts of Northeast Hurricanes on Energy Infrastructure 50.

Vadlamudi, V.V., Hamon, C., Gjerde, O., Kjolle, G., Perkin, S., 2016. On Improving Data and Models on Corrective Control Failures for Use in Probabilistic Reliability Management, in: 2016 International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2016 - Proceedings. https://doi.org/10.1109/PMAPS.2016.7764089

Vanschoren, J., 2018. Meta-Learning : A Survey 1–29.

Vefsnmo, H., Kjølle, G., Jakobsen, S.H., Ciapessoni, E., Cirio, D., Pitto, A., 2015. Risk Assessment Tool for Operation : from Threat Models to Risk Indicators, in: PowerTech Eindhoven 2015. IEEE Press, p. 6. https://doi.org/10.1109/PTC.2015.7232557

Villa, V., Paltrinieri, N., Khan, F., Cozzani, V., 2016. Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. Saf. Sci. 89, 77–93. https://doi.org/10.1016/J.SSCI.2016.06.002

Wanik, D.W., Anagnostou, E.N., Hartman, B.M., Frediani, M.E.B., Astitha, M., 2015. Storm outage modeling for an electric distribution network in Northeastern USA. Nat. Hazards 79, 1359–1384. https://doi.org/10.1007/s11069-015-1908-2

Wanik, D.W., Parent, J.R., Anagnostou, E.N., Hartman, B.M., 2017. Using vegetation management and LiDAR-derived tree height data to improve outage predictions for electric utilities. Electr. Power Syst. Res. 146, 236–245. https://doi.org/10.1016/j.epsr.2017.01.039

Yang, X., Haugen, S., 2015. Classification of risk to support decision-making in hazardous processes. Saf. Sci. 80, 115–126. https://doi.org/10.1016/j.ssci.2015.07.011

Yang, X., Haugen, S., Paltrinieri, N., 2018. Clarifying the concept of operational risk assessment in the oil and gas industry. Saf. Sci. 108, 259–268. https://doi.org/10.1016/j.ssci.2017.12.019

Yu, W., Pollitt, M.G., 2009. Does Liberalisation cause more electricity blackouts ? Evidence from a global study of newspaper reports study of newspaper reports. https://doi.org/https://doi.org/10.17863/CAM.5280

Appendices

Appendix A - Listing of accessible Tools used during the Ph.D..

A. *Appendix A - Listing of accessible Tools used during the Ph.D.*

Table A – Reporting of the main important accessible tools used during the Ph.D.

N°	Tool name	Nature of the tool	Provided support	Address
1	CloudCompare	Open-source 3D point cloud and mesh processing software	Enabled analysis and conversion of 3D point clouds and meshes	www.cloudcompare.org
2	Colmap	Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline	Used to generate photogrammetry point clouds	https://colmap.github.io/
3	Converter	Geographical unit converter	Used to compare coordinates in different projections.	rcn.montana.edu/Resources/Converter.aspx
4	Converter	UTC/GPS Time Converter	Enabled to find the UTC equivalent of GPS time	https://www.gw-openscience.org/gps/
5	Displaz	Free LiDAR point cloud viewer	Enabled the visualisation of LiDAR point clouds	c42f.github.io/displaz
6	ExifTool	Meta-information editor	Enabled to read meta information in images	exiftool.org
7	LASTools	Free LiDAR toolbox	Enabled metadata extraction of point clouds using the "lasinfo" tool	rapidlasso.com/lastools
8	PyCharm	Integrated Development Environment	Used for computer programming	https://www.jetbrains.com/pycharm/
9	QGIS	Open-source GIS application	Enabled visualisation of maps	www.qgis.org
10	Sublime Text	Source code editor	Enabled easy looks into code	https://www.sublimetext.com/

Part II – Articles

Article I - Risk of crack formation in power grid wooden poles and relationship with meteorological conditions : a Norwegian case study

Pacevicius, M., Roverso, D., Salvo Rossi, P., Paltrinieri, N., 2018b. **Risk of crack formation in power grid wooden poles and relationship with meteorological conditions : a Norwegian case study**, in: European Safety and Reliability Conference - ESREL 2018. Trondheim, p. 7.

Risk of crack formation in power grid wooden poles and relationship with meteorological conditions: a Norwegian case study

Michael Pacevicius
eSmart Systems, Halden, Norway

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, Trondheim, Norway

Davide Roverso
eSmart Systems, Halden, Norway

Pierluigi Salvo Rossi
Kongsberg Digital, Trondheim, Norway

Nicola Paltrinieri
Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, Trondheim, Norway

ABSTRACT: Predicting the occurrence of failures in power grids through specific outage risk predictors is a primary concern for utilities nowadays. Wooden poles represent core items to focus on in this process. Millions of them are used worldwide and they are all subject to the risk of crack formation. Analyzing the evolution of pole cracks is particularly relevant in reliability analyses of power grids for two main reasons. First: the cracks might highlight previously unconsidered or changing factors, such as unusual local weather conditions (e.g. overload of ice and/or wind). Second: as cracks provide an access for external threats (e.g. humidity, fungi, insects) to potentially non-treated internal parts of the poles, they might in turn accelerate the occurrence of further failures. Evaluating the role of crack formation is thus essential for estimating the risk of outages in power grids. As climatic variations are known to be among the most influencing factors in the initiation and propagation of cracks in wooden poles, we address this topic by suggesting a method combining open-access weather-data sources with information provided by new technologies, such as drones. We first highlight the influence of climatic factors on the reliability of wooden poles by reviewing studies describing the physical properties of wood. We then focus our research on a Norwegian case study and show how we can combine up to 60 years of meteorological information with the information provided by 17,352 geo-localized aerial pictures of cracked and non-cracked wooden utility poles. We finally discuss the way an indicator constructed on this combination can be used to predict the formation of cracks and optimize the allocation of decision-maker resources for inspection procedures.

1 INTRODUCTION

The modernization of the society has led to a global increase of power consumption over the last 50 years (Refsnæs, Rolfseng, Solvang, & Heggset, 2006; Shiu & Lam, 2004; Yoo & Kwak, 2010). As numerous businesses, public infrastructures and private households rely on the provision of power for their daily tasks, there is a need for companies in charge of the power supply to maximize their capacity and reliability in delivering power.

Predicting outage risks and avoiding downtime is crucial to ensure customer satisfaction. Moreover, anticipating unwanted events directly enables power utilities to significantly reduce losses and costs. Finally, it also enables them to optimize resource allocations for the inspection of their infrastructures after natural disasters (e.g. storms, flooding) or during scheduled maintenance procedures.

Ensuring this quality of service requires utilities to use reliable components, from the power source, through the transmission lines and to the consumption nodes. Wooden poles are widely used for the distribution part of the power grid (from regional substations to local substations and from local substations to end-users) (Eurelectric, 2010).

Identifying the principal factors responsible for the apparition of cracks in wooden poles represents thus a main objective for predicting their failures. For this purpose, we suggest a method enabling to evaluate the effects of potential predictors. The contribution identifies the way forward for this research topic and presents preliminary findings, representing the basis for future research.

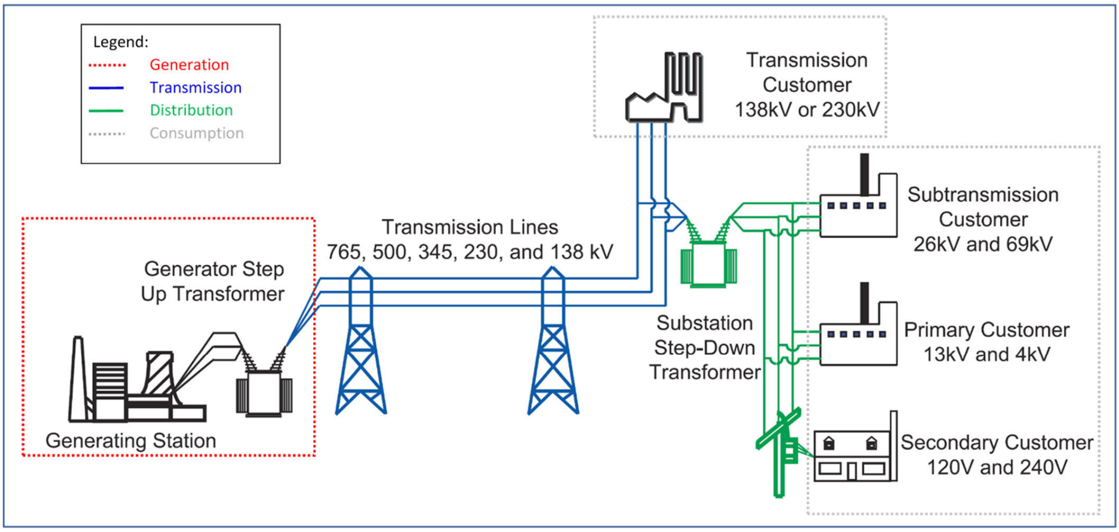


Figure 1. Outline of the transmission and distribution of power in a power grid, going from the production sites to the consumption nodes. Adapted from (U.S.-Canada Power System Outage Task Force, 2004).

The rest of the paper is constructed as follows. Section 2 provides an overview on wooden poles characteristics and failures. Section 3 mentions various studies summarizing the main properties of wood on microscopic level. On this basis, it highlights the influence climatic variations can have on the physical structure of wooden poles. It furthermore shows how the variations can affect the reliability of the pole and thus of the transmission line. Section 4 describes the strategy applied to provide values of a crack-apparition likelihood using a Norwegian case study. It explains the choices made in the selection of the different datasets and the methods used to acquire them. Section 5 discusses the pros and the cons of the method used and shortly describes plans for future research. The last section finally concludes our work by summarizing and suggesting additional research possibilities.

2 WOODEN POLES CHARACTERISTICS AND FAILURES

Figure 1 shows schematically how power is delivered from a generating station, through transmission and distribution lines (respectively maintained by Transmission System Operators (TSO) and Distribution System Operators (DSO)), to different categories of end customers. Wooden utility poles used in the power grid exist in different shapes and configurations, depending on the physical requirements of the power lines, on the geographical conformation of their location, and on their position in the transmission or distribution line (see figures 2-4 as illustrations).

Figure 2. First example of the shape of a wooden utility pole

Figure 3. Second example of the shape of a wooden utility pole

Figure 4. Third example of the shape of a wooden utility pole

Despite the variety of the existing shapes and configurations, the number of elements basically composing an electrical pole is relatively limited. A wooden utility pole is generally composed of one or more wooden poles, one or more cross-arms and multiple insulators responsible for the junction between the electrical cables and the pole. Figure 5 schematizes this assembling.

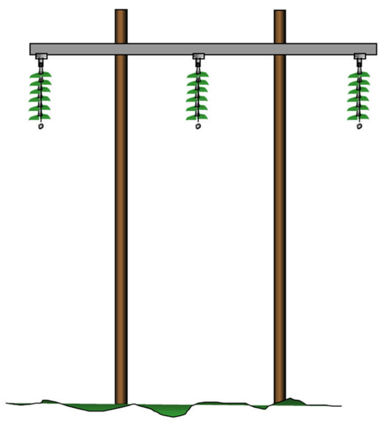


Figure 5. Basic components of a wooden utility pole: poles (brown), cross-arm (grey) and insulators (green) (Refsnæs, 2008)

Using wooden utility poles has multiple advantages in comparison to concrete or steel utility poles (Bolin & Smith, 2011; SEMCO, 1992; Stewart, 1996)

- They are lighter and easier to transport on mountainous fields.
- They do not require earthing, which makes them interesting when lightning occur.
- They are easy to produce in wooded areas (e.g. Canada, Norway).
- They generally have a reduced environmental impact.
- They have interesting lifetimes, possibly going up to 75 years in favorable conditions.

Identifying the main threats for wooden utility poles enables to look for root causes of failures. This gives the possibility to estimate their effective remaining lifetime and optimize their replacement before any outage.

In their review on power line inspection procedures, Nguyen et al. (Nguyen, Jenssen, & Roverso, 2018) summarize some of the main common faults of power line components. They identify the apparition of cracks in the wooden poles as being one of the main failure to identify during visual inspection procedures. An additional review of the literature shows that there is need for inspection protocols enabling to recognize and assess cracks in timber structures in general (Dubois, Chazal, & Petit, 2002; Riahi, Moutou Pitti, Dubois, & Chateauneuf, 2016) and in wooden poles in particular (Morrell, 2012) .

Identifying cracks is fundamental for two main reasons:

- First, as “stresses perpendicular to grain induce cracks which propagate longitudinally” (Coureau & Morel, 2005), we can consider multiple apparitions of significant cracks as being indicators of the presence of stress factors. This can for example suggest the existence of a localized area subject to harsher weather conditions (e.g. overload of ice and/or wind) (Wong & Miller, 2010) and prompt deepened analysis of the concerned region.
- Second, as cracks provide an access for external threats (e.g. fungi, insects, humidity) to potentially non-treated internal parts of the poles, their existence might accelerate the apparition of decay (Morrell, 2012; Refsnæs et al., 2006; SEMCO, 1992). This permanently alters the structural resistance of the pole and considerably increases its probability of failure.

3 WOOD PROPERTIES AND POTENTIAL INFLUENCE OF CLIMATIC VARIATIONS ON CRACK APPARITION

The theory of fracture mechanics has mainly been developed since the first half of the 20th century. Initiated by A. A. Griffith in 1920 (Griffith, 1921), it has then been popularized by G.R. Irwin in 1958 (Irwin, 1958) and is since being widely used to analyze the origins and consequences of crack apparition in physical objects. Focusing on the microscopic level, it enables to provide models describing the “mechanical behavior of cracked materials subjected to applied load” (Perez, 2017) .

Multiple studies use this theory as a basis for the evaluation of crack growth in wooden structures (Barrett, Haigh, & Lovegrove, 1981; Coureau & Morel, 2005; Dubois et al., 2002; Riahi et al., 2016). A characterization of the structure is initially made on microscopic level to understand how wood behaves when it is subject to a modification of its external environment (load variation, climatic variation, etc.). Figure 6 shows the structure on microscopic level of a typical softwood. It highlights the anisotropic characteristic of wood and intuitively shows that cracks are more probable to occur parallel to the direction of growth of a three (longitudinal direction).

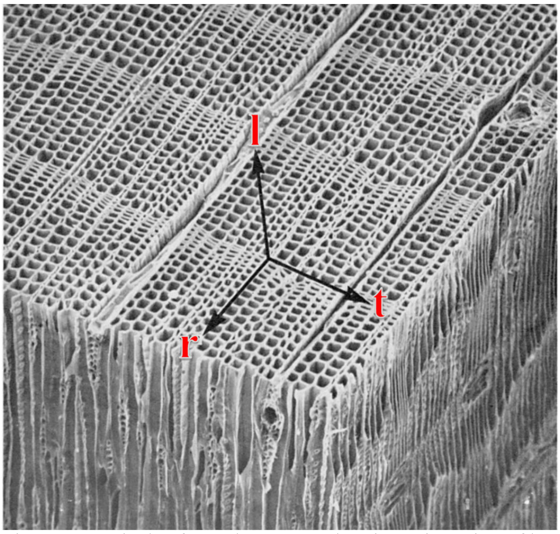


Figure 6. Typical softwood structure showing orientation of longitudinal (l), radial (r) and tangential (t) directions (Barrett et al., 1981).

Wood being furthermore a viscoelastic material, its physical properties (e.g. modulus of elasticity, volume) are directly influenced by their environment. This is due to the hygroscopic behavior of wood (i.e. tendency to absorb humidity) and implies that physical properties of wood are highly sensitive to the meteorological properties of its surrounding (especially temperature and humidity) (Chaplain & Valentin, 2010; Hamdi, Moutou Pitti, & Saifouni, 2017; Lamy, 2016; Morrell, 2012; Refsnæs et al., 2006; Saifouni, 2014; Thybring, Lindegaard, & Morsing, 2009).

Because of the former functionalities of their cells during their living period and because of the variations in their environment during their growth, mechanical properties of timber-based structures can furthermore be locally modified. This includes structure modifications due to natural defects such as knots, rotten knots holes or cracks due to freezing lifeblood. Combined with the application of external loads (e.g. wind, ice on the wires in the case of wooden poles) and the modification of its internal structure due to temperature and humidity variations, there is a fertile ground for the apparition of cracks.

4 DATA ACQUISITION AND PREDICTION METHODS

Utility companies in Norway use over 3.5 million wooden poles in their power grids to support over 25,400 km of electrical overhead lines (Eurelectric, 2010; Refsnæs et al., 2006). The Norwegian IT company eSmart Systems[1] is specialized in digital intelligence and uses artificial intelligence to support Statnett, Norway's TSO, as well as some of the main Norwegian DSOs (e.g., Lyse Elnett, Ringeriks-Kraft Nett, Troms Kraft Nett, Hafslund Nett). In particular, the algorithms used by eSmart Systems automatically identify specific objects and recognize pre-defined faults, such as cracks on wooden poles (see figure 7 as an illustration). This enabled us to access a database of 17,352 geo-localized aerial pictures of wooden utility poles, from which 5383 are classified as cracked.

Figure 7. Wooden pole where a crack has been localized on the mast (see rectangle).

In most of the cases, two to three pictures of a unique utility pole were taken from different angles. This was done to ensure having accurate information for each of the observed poles without suffering from hidden information. We merged this information with the exact geographical coordinates of the electric poles, made available by the Norwegian Water Resources and Energy Directorate (NVE)[2]. We could thus analyze a dataset of 7653 geo-localized wooden utility poles, either classified as cracked or not.

In parallel, seNorge[3] (created in collaboration between the NVE, the Norwegian Meteorological Institute[4] and the Norwegian Mapping Authority[5]) enables us to access daily observed (or interpolated) records of the climatic conditions in Norway. Especially, it enables us to access temperature and precipitation measures going as far back as 1957.

[1] eSmart Systems: www.esmartsystems.com
[2] NVE: www.nve.no
[3] seNorge: www.senorge.no
[4] Norwegian Meteorological Institute: www.met.no
[5] Norwegian Mapping Authority: www.kartverket.no

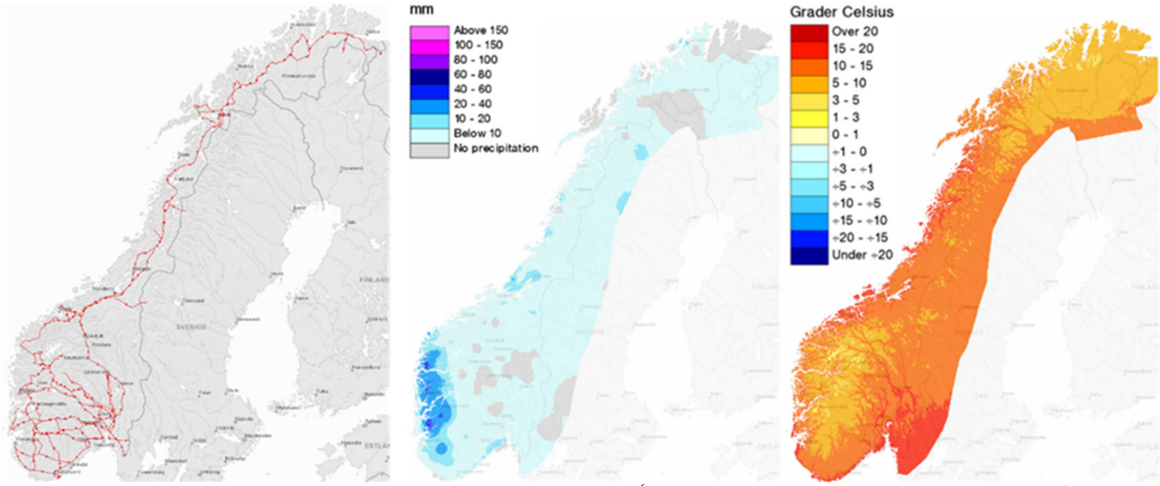


Figure 8. From left to right: main axes of the Norwegian electrical grid[6]; map of the precipitation in Norway on the 1st of August 2017[7]; map of the temperatures in Norway on the 1st of August 2017[7].

Figure 8 illustrates the type of information made available by NVE and seNorge. Using a scroll up/down feature of the websites, it is possible to move from a global and national overview up to a specific geo-localized point (in our case, the localization of the wooden utility poles).

Different approaches are considered in our work. The purpose is to create an indicator for the likelihood of crack apparition on wooden poles.

In order to benefit from the high granularity offered by the webservices used, we plan to use daily records of temperature and precipitation as potential predictors for a binary classification problem (labeling as cracked or not-cracked). Predictive features can be designed, that summarize at different granularities the daily weather data and extract relevant indicators that correlate with crack appearance. Considering an extreme reduction, we can for example summarize the intensity of the meteorological variation on a localized point into, e.g. a temperature coefficient and a precipitation coefficient. This would lead to a method using only two predictors when focusing on this classification problem.

Equation (1) provides an example of the type of coefficient c that can be used when focusing on a specific pole.

$$c = \sum_{i=2}^{n} \frac{|X_i - X_{i-1}|}{X_{max} - X_{min}} \tag{1}$$

Where n is the number of daily records since the installation of the wooden pole observed; i the enumeration index; X_i the value of the meteorological phenomenon observed on the specified location on day i (here in millimeters or in degrees Celsius); X_{i-1} the record of the same phenomenon on the same location on the previous day; X_{max} (resp. X_{min}) the maximum (resp. minimum) value of the observed phenomenon that has been recorded over the entire timestamp of observation on the specified location.

Alternatively, predictive features can be automatically learned from the raw temperature and precipitation time series using deep learning techniques. Such techniques, belonging to the class of artificial intelligence methods (and more especially, to the class of machine learning methods) are based on recursive analyses of data over time and/or over space, from which they identify and highlight step by step the most relevant characteristics.

High temperatures favor the proliferation of fungus, which weakens the structure of the wood. Furthermore, high humidity levels on extended periods might soften the wood and make it more sensitive to sudden external loads (e.g. wind or ice rain). Finally, the intrinsic properties of wood lead it to easily accept slow variation of external loads and environmental conditions but make it particularly sensitive to sudden variations. These approaches will thus enable us to identify meteorological patterns favoring the apparition of cracks, as well as located regions where the likelihood of crack apparition will be higher.

An increase in the period of exposition to external factors leads to a rise of the probability of crack apparition. This implies that the age of the poles plays a big role in the suggested methods. However, part of this information might be missing. In such a case, we could consider a generic day of installation depending on the period of installation of the power line in the observed region.

[6] https://temakart.nve.no/link/?link=nettanlegg
[7] http://www.senorge.no/index.html?p=senorgeny&st=weather

5 DISCUSSION

The suggested methods enable to evaluate the role that temperature variations and precipitations have on the formation of cracks on wooden poles. These methods have the advantage to be flexible and easily integrated when accessing additional data sources, such as daily records of wind intensity and direction, humidity variations, clouds presence, etc. They are nevertheless highly dependent on two main facts:

- First, the initial classification of the poles as cracked or not. This is an important topic as the size of the cracks directly affects its detection by the algorithm used to classify the poles. There is thus a need for utility companies to define what should be considered as a problematic crack or not.
- Second, the information initially available on the poles themselves (e.g. age, maintenance tasks carried out). This information might be difficult to access because not necessarily well reported in the first phases of the grid installation.

Despite using relatively simple techniques and being highly dependent on initial parameters, the proposed methods represent a first approach in the analysis and handling of cracks in wooden poles. This information may in turn be useful for decision makers in the prioritization of additional inspection procedures and future maintenance tasks.

It is to mention that our paper only highlights preliminary results of an ongoing research, as the described methods have not yet been fully applied. Further work will thus focus on the extensive application and validation of these approaches and provide an in-depth analysis of the phenomenon of crack apparition on wooden poles by using additional real data from the Norwegian network.

6 CONCLUSION

Our paper highlighted the importance for utilities of early detection and analysis of cracks on wooden poles. We summarized how environmental conditions can directly affect the physical properties of wood and thus favor or limit the apparition of cracks on wooden poles. In order to better understand and predict their occurrence, we then suggested two approaches using pre-classified and geo-localized aerial pictures of cracked and non-cracked poles in combination with up to 60 years of meteorological measurements. Further, we saw that, despite being highly dependent on initial information, our approach might provide useful information for the generation of maintenance policies. This approach might finally be a good starting point for researchers wanting to combine fields of expertise such as structural study of wood on microscopic level and crack detection methods using image analysis.

7 REFERENCES

Barrett, J. D., Haigh, I. P., & Lovegrove, J. M. (1981). Fracture Mechanics and the Design of Wood Structures. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *299*(1446), 217–226. https://doi.org/10.1098/rsta.1981.0020

Bolin, C. A., & Smith, S. T. (2011). Life cycle assessment of pentachlorophenol-treated wooden utility poles with comparisons to steel and concrete utility poles. *Renewable and Sustainable Energy Reviews*, *15*(5), 2475–2486. https://doi.org/10.1016/j.rser.2011.01.019

Chaplain, M., & Valentin, G. (2010). Effects of Relative Humidity Conditions on Crack Propagation in Timber: Experiments and Modelling. In *World Conf. on Timber Engineering* (pp. 1–8). Retrieved from http://support.sbcindustry.com/Archive/2010/june/Paper_438.pdf?PHPSESSID=ju29kfh90oviu5o371pv47cgf3

Coureau, J. L., & Morel, S. (2005). Non-Linear Fracture Mechanics Applied To Wood In Mode I. In *ICF11* (pp. 1–6). Italy. Retrieved from http://www.gruppofrattura.it/ocs/index.php/ICF/ICF11/paper/viewFile/10698/10044

Dubois, F., Chazal, C., & Petit, C. (2002). Viscoelastic crack growth process in wood timbers: An approach by the finite element method for mode I fracture. *International Journal of Fracture*, *113*(4), 367–388. https://doi.org/10.1023/A:1014203405764

Eurelectric. (2010). *EURELECTRIC's views on the use of creosote for impregnation of wooden poles in electricity networks*. Brussels, Belgium. Retrieved from http://www.eurelectric.org/media/44303/eurelectric_comments_on_creosote_2010-11-16-2010-030-1024-01-e.pdf

Griffith, A. A. (1921). The phenomena of rupture and flow in solids. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *221*, 163–198. Retrieved from http://www.jstor.org/stable/91192

Hamdi, S. E., Moutou Pitti, R., & Saifouni, O. (2017). Moisture driven failure monitoring in wood material: numerical analysis based on viscoelastic crack growth approach. In *CompWood 2017 – ECCOMAS Thematic Conference on Computational Methods in Wood Mechanics – from Material Properties to Timber* (pp. 187–198). Retrieved from https://www.researchgate.net/profile/Rostand_Pitti/publication/317329144_Moisture_driven_failure_monitoring_in_wood_material_numerical_analysis_based_on_viscoelastic_crack_growth_approach/links/59328dac0f7e9beee791a678/Moisture-driven-failure-monitoring-in-wood-material-numerical-analysis-based-on-viscoelastic-crack-growth-approach.pdf

Irwin, G. R. (1958). Fracture. In F. S. (Ed.), *Elasticity and Plasticity / Elastizität und Plastizität* (Vol. 3 / 6, pp. 551–590). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-45887-3_5

Lamy, F. (2016). *Crack analysis in wood under mechanical and climatic loadings : Contribution of Acoustic Emission*. Université de Limoges. Retrieved from https://tel.archives-ouvertes.fr/tel-01364070

Morrell, J. J. (2012). *Wood Pole Maintenance Manual: 2012 Edition*. Oregon State University. Forest Research Laboratory. https://doi.org/http://ir.library.oregonstate.edu/concern/technical_reports/ft848r69b

Nguyen, N. Van, Jenssen, R., & Roverso, D. (2018). Automatic Autonomous Vision-based Power Line Inspection : A Review of Current Status and the Potential Role of Deep Learning. *International Journal of Electrical Power & Energy Systems*.

Perez, N. (2017). *Fracture Mechanics*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24999-5

Refsnæs, S. (2008). *Lineoppheng*. Retrieved from http://docplayer.me/36970970-Lineoppheng-sintef-energiforskning-as.html

Refsnæs, S., Rolfseng, L., Solvang, E., & Heggset, J. (2006). Timing of wood pole replacement based on lifetime estimation. In *9th International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2006* (pp. 1–8). https://doi.org/10.1109/PMAPS.2006.360286

Riahi, H., Moutou Pitti, R., Dubois, F., & Chateauneuf, A. (2016). Mixed-mode fracture analysis combining mechanical, thermal and hydrological effects in an isotropic and orthotropic material by means of invariant integrals. *Theoretical and Applied Fracture Mechanics*, *85*, 424–434. https://doi.org/10.1016/j.tafmec.2016.06.002

Saifouni, O. (2014). *Modeling of rheological effects in materials : application to the mecanosorptive behaviour of wood*. Universit_e Blaise Pascal - Clermont-Ferrand II. Retrieved from https://tel.archives-ouvertes.fr/tel-01069026/

SEMCO. (1992). *Wood pole maintenance. Bureau of Reclamation, Facilities Instructions, Standards, and Techniques* (Vol. 4–6). Retrieved from https://www.usbr.gov/power/data/fist/fist_vol_4/vol4-6.pdf

Shiu, A., & Lam, P.-L. (2004). Electricity consumption and economic growth in China. *Energy Policy*, *32*(1), 47–54. https://doi.org/10.1016/S0301-4215(02)00250-1

Stewart, A. H. (1996). *How long do wood poles last?* Fort Collins. Retrieved from http://www.americanpoleandtimber.com/wp-content/uploads/how-long-do-wood-poles-last.pdf

Thybring, E. E., Lindegaard, B., & Morsing, N. (2009). Service Life Prediction of Wood Claddings by in-situ Measurement of Wood Moisture Content: Status after 5 years of Outdoor Exposure. In *40th Annual Meeting of the International Research Group on Wood Protection*. Beijing, China. Retrieved from https://www.researchgate.net/profile/Emil_Thybring/publication/262258436_Service_life_prediction_of_wood_claddings_by_in-situ_measurement_of_wood_moisture_content_status_after_5_years_of_outdoor_exposure/links/00b7d5372224cd1f31000000/Service-life-prediction-of-wood-claddings-by-in-situ-measurement-of-wood-moisture-content-status-after-5-years-of-outdoor-exposure.pdf

U.S.-Canada Power System Outage Task Force. (2004). *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations* (Vol. 40). Washington, DC, US. Retrieved from https://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/BlackoutFinal-Web.pdf

Wong, J. C., & Miller, M. D. (2010). *Guidelines for Electrical Transmission Line Structural Loading*. Reston, Virginia: American Society of Civil Engineers. https://doi.org/10.1061/9780784410356

Yoo, S. H., & Kwak, S. Y. (2010). Electricity consumption and economic growth in seven South American countries. *Energy Policy*, *38*(1), 181–188. https://doi.org/10.1016/j.enpol.2009.09.003

Article II - Smart Grids : Challenges of Processing Heterogeneous Data for Risk Assessment

Pacevicius, M., Roverso, D., Salvo Rossi, P., Paltrinieri, N., 2018a. **Smart Grids : Challenges of Processing Heterogeneous Data for Risk Assessment**, in: Probabilistic Safety Assesment and Management - PSAM 14. Los Angeles, p. 11.

Smart Grids: Challenges of Processing Heterogeneous Data for Risk Assessment

Michael Pacevicius[*a, b], Davide Roverso[a], Pierluigi Salvo Rossi[c] and Nicola Paltrinieri[b]

[a] eSmart Systems, Halden, Norway

[b] Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, Trondheim, Norway

[c] Kongsberg Digital, Trondheim, Norway

Abstract: Recent advances in IT-related fields are opening up a broad range of novel applications. This is especially true in the energy sector, where Smart Grid solutions are offering new opportunities for the monitoring of power transmission and distribution in electrical grids. However, optimal use of potentially accessible data sources is challenging, and most of the current Smart Grid projects continue to exhibit suboptimal utilization of heterogeneous information. This situation is also faced when it comes to the assessment of risks associated to operation of electricity transmission and distribution networks. As a consequence, current management systems fail to provide accurate estimations of risk levels in real-world situations. Our paper addresses this issue and contributes to the identification of possible solutions. The paper identifies a number of heterogeneous data sources which could be relevant for risk assessment, but which are currently not fully exploited. Furthermore, the paper points to valuable relations existing across these data sources, that promote a better understanding of real-world situations and empower a more accurate analysis of well-known or newly identified risks within the framework of Smart Grids.

Keywords: Risk assessment, Smart Grids, Heterogeneous Datasets, Data Acquisition, Link Identification.

1. INTRODUCTION

Digitalization offers numerous advantages thanks to the progress made in the IT field (e.g., increase of memory capacities, increase of processing capacities, development of cross-technologies collaboration platforms). For instance, it enables to interconnect devices, to access them through multiple communication networks in real time, and to handle the large quantity of data they generate and transmit. “Smart Grids”, which represent an example of this idea application in the energy sector, offer interesting perspectives for the management of power grids.

There may be several benefits of Smart Grids, such as: higher demand response with minimized costs, reduction of the environmental impact and integration of renewable energy resources, and resilience to disturbances as well as electrical stability in the grid.

Smart Grids may also enable system operators to reduce outage risks by getting access to previously unconsidered data, ranging from weather forecasts to social network data. Combining outage reports with weather reports could for example improve risk monitoring in regions with harsh climatic conditions. This approach has been previously explored in some projects [1], [2], whose focus has especially been set on the impact of climatic conditions. Unfortunately, current researches do rarely fully exploit the real precision degree offered by todays IT solutions and services, and many projects still use averaged data and meta-data for their analyses. Similarly, the full range of various data sources, from which numerous datasets are made accessible thanks to open-access policies (e.g., *OpenAire*, *ENTSO-E Transparency Platform*, *U.S. Open Data platform*), is far from being optimally exploited, especially when it comes to risk assessment.

[*] michael.pacevicius@esmartsystems.com

This challenge is demonstrated by three main arguments:

1. The available frameworks and standards from industrial risk management (e.g., *CSA Q850-97*, *ISO 31000:2009*, *NORSOK Z-013*) are generally overlooked when it comes to the study of Smart Grids. Furthermore, only a few studies propose effective solutions enabling, based on internal as well as external factors, dynamic updates of a risk management framework [3]. This reduces the advantages that can be derived from actual digitalization and hinders the exploitation of the real-time feature offered by Smart Grids.

2. Although Probabilistic Safety Assessment (PSA) approaches advocate for the constitution of cross-disciplinary expert teams (which may include systems analysts, PSA specialists, operators and operational analysts, data scientists and human factor analysts [4]), risk analysis is mainly based on collections of operating and maintenance data, without taking advantage of additional accessible data. In fact, enlarging the horizon of the observations, in association with new data sources, could allow controllers to detect, observe and potentially predict slow, long-term and non-trivial phenomena (e.g., mechanical fatigue, corrosion, dust accumulation) increasing the failure probabilities.

3. The lack of cross-disciplines experts hinders decision makers from identifying relevant links between data sources, compromising the recognition of efficient combinations of data sources.

There is thus need for new methods enabling continuous and effective integration of heterogeneous data for accurate risk assessment predictions. For this reason, this paper focuses on the first phases of Smart Grid dynamic risk management: collection and combination of relevant datasets. It is constructed as follows. Following the present introduction, the second section reviews the main concepts defining the Smart Grid technology. It describes Smart Grids, mentions the role of the mains organizations, committees and work groups focusing on the topic, and lists the main challenges faced. The third section focuses on the risk dimension in the context of Smart Grids. It shows how the utilization of Smart Grid technologies affects the risk level in the context of power grid management. It continues by pointing out the role the treatment of heterogeneous datasets can have in this context and underscores the importance of the *Common Information Model* (CIM). It finally brings forward work initiated by different research groups on the use of heterogeneous data for risk assessment in power grids. Section 4 describes the investigation procedure. It reminds the main objectives of our work and shortly describes the principles followed to reach our goal. Section 5 describes the results of this paper. It reports a list of diverse data sources newly identified as being relevant for the enhancement of risk assessment, and provides knowledge for merging this information. Section 6 (discussion) highlights the way forward for risk reduction in Smart Grids, but also provides understanding on what remains to be done in the field. The last section (conclusion) shortly reviews the paper and finally concludes our work.

2. OVERVIEW OF SMART GRIDS

The emergence of the first automatic meters and the broadening of SCADA (Supervisory Control And Data Acquisition) technologies in power grids represent the first examples of Smart Grid concepts. Since the creation of the *European Smart Grids Technology Platform* by the European Commission [5], Smart Grids have gained increasing attention and represent nowadays a well-known topic.

Different definitions are given to Smart Grids by scientific literature and industry [6]–[9]. However, they can be summarized as follows:

The basic concept of a Smart Grid is to optimize the production, transmission and consumption of power and information between the different elements (devices & actors) involved in it. For this purpose, it makes use of the progresses made in the information and communication technology (ICT) sector, as well as those made in the industry of electronic components. Thanks to these advances, it is possible to integrate modern technologies (e.g., photovoltaic panels, batteries, smart meters and other types of devices made accessible through the use of connected sensors) into an existing power grid. This eases and automates management for a number of tasks, such as production coordination, energy distribution and power consumption.

A broadening range of communication channels (internet, satellite, etc.) enable real-time access to the large amount of generated data. The high-resolution picture of interconnected devices that is given empowers improved analyses, finer forecasts of future consumption trends and better predictions of potential outages.

The utilization of low-carbon alternatives for power production (solar panels, wind turbines, water turbines, etc.) and the expanded utilization of batteries have led to a shift from a centralized to a distributed repartition of the production sites. The gained autonomy has, in combination with the intelligence integrated in the power grid, boosted the emergence of micro-grids, which are capable to be islanded and reconnected ("split & merge") to external parts of the network without major disturbances.

Micro-grids present multiple advantages: (1) they provide flexibility in case of outages and enable to maintain power supply during programmed maintenance tasks; (2) the technologies utilized can rapidly be implemented and assimilated; (3) these same technologies can be used for shaving peaks under normal circumstances by compensating the needs of one another. In the same idea, Smart Grids have also the capacity to expand and assimilate new items without hardly impacting the general stability of the grid ("integrate as you grow"). This is enabled by the flexibility acquired, which has led to a dynamization and a decentralization of the decision making. As a consequence, end-users are more involved and share henceforth management responsibilities with power companies, to which they are now capable to smartly and easily provide positive and negatives feedbacks. End-users do not anymore only represent passive consumption nodes, but they become active actors and turn themselves into deciding prosumers (producer-consumer). This completely disrupts the flow management, especially because old infrastructures and protection systems – which were originally conceived to be unidirectional – have now to handle reverse flows of energy. In addition to the challenging multi-directional aspect of power management, multiple other stumbling blocks can also be enumerated when it comes to Smart Grids. A list of the most obvious ones is given in table 1.

Table 1: Most Important Challenges in Smart Grids

N°	**Stumbling Blocks**
1	The collaboration between the main stakeholders and the interconnection of multiple platforms/technologies can be hindered for compatibility reasons or because of conflicts of interest.
2	Although the integration of new technologies enables to compensate some of the fluctuations formerly existing in the power grid, such technologies can also themselves be responsible for new type of peaks. High spontaneous variations of power flows (e.g. because of clouds passing over a solar panel or sudden reduction of the wind intensity in a windfarm) might lead to failures in an aging infrastructure not conceived for such drastic fluctuations.
3	The data made accessible to the customer might lead them to change their habits into unknown ways, which can imply unexpected variation and calls for new profiling analyses.
4	New types of variations might reduce the efficacity of older fault detection methods, localization methods and temporary proxy methods, which are based on assumptions of more stable behaviors in the grid.
5	Smart Grids projects encounter classical *big data* challenges [10] and considerable analytical efforts need to be done in order to transform *data* into *information*. The lack of adapted tools often leads to the ignoring of data because no proper method could enable them to be integrated [11].
6	Real-time or near-real-time access and analysis of data still concerns only a restricted number of information sources. Much information remains post-transmitted and post-processed with more or less high latency (from hours [e.g. with nightly updates of servers] to years [e.g. with the inspection for maintenance of distant power lines]).
7	Optimization techniques for the management of Smart Grids still need to be improved, especially because it is challenging to handle the exponentially growing number of assets integrated in the grid. Finding an optimum between production, transportation, storage and consumption of power in a distributed network is known to be a complicated task.
8	The concept of Smart Grids is closely related to the concept of smart cities, where the management and optimization of water grids, gas pipelines and heating/cooling networks is similarly challenging. This might reinforce every one of the previously highlighted challenges.

To address these numerous challenges, different work groups and committees have been constituted all over the world. Those can exist on global level (North America[1], European Union[2], Asia-Pacific Economic Cooperation[3] or even on larger level[4,5]), but many countries (Korea, Japan, Norway, France, Germany, etc.) do also have national energy regulation committees and Smart Grids associations supporting research in the field. These different organizations contribute to the spreading of knowledge around the topic of Smart Grids, by providing technical reports, protocols, knowledge about standards and by favoring open access demarches and cross-vendor-compatible solutions. The *Smart Electric Power Alliance* (SEPA) provides for example an interesting and regularly updated Catalog of Standards (CoS) on their webpage[6], which enables one to rapidly be aware of the most relevant standards in the field of Smart Grids. The listed standards emanate from standardization organizations and research institutions (e.g., the *Electric Power Research Institute* (EPRI), the *National Institute of Standards and Technology* (NIST), the *International Electrotechnical Commission* (IEC), the *European Committee for Standardization* (CEN), the *European Committee for Electrotechnical Standardization (*CENELEC), the *European Tele-communications Standards Institute* (ETSI)), which permanently provide updates for old and new standards.

[1] North American Electric Reliability corporation (NERC) Smart Grid Task Force
[2] EU Smart Grids Task Force
[3] Energy Smart Communities Initiative (ESCI)
[4] International Smart Grid Action Network (ISGAN)
[5] Global Smart Grid Federation (GSGF)
[6] https://sepapower.org/knowledge/catalog-of-standards

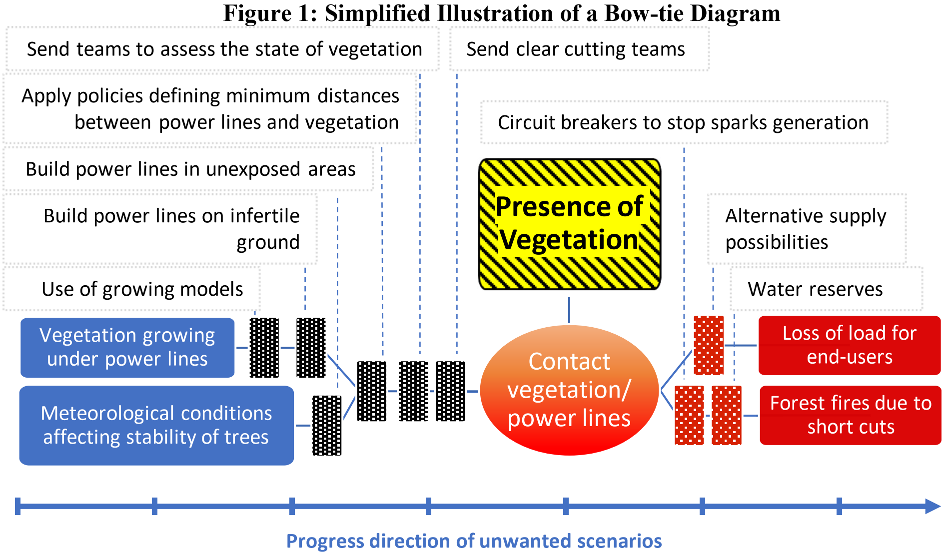
3. RELEVANT RISKS & RELATED INFORMATION

3.1. Risks in the Power Sector

Kaplan and *Garrick* [12] define the general notion of risk by characterizing the outcome of a specific action. More especially, they consider risk as a triplet of (1) a potential unwanted event, (2) its likelihood of occurrence and (3) the consequences of this event happening.

The veracity of the three dimensions depicting risk is dependent on the level of knowledge of a specific situation one may have at a specific moment. This defines the notion of uncertainty, which highlights the incapacity to perfectly characterize a scenario (some events are not anticipated; some probabilities are inadequately trusted; and some consequences are not foreseen). The process of risk assessment aims to reduce this uncertainty by periodically reviewing risky situations and suggesting barriers. Therefore, it uses tools such as the bow-tie diagram [13], which allow for identification and characterization of possible scenarios associated to a hazard [3]. Figure 1 schematically represents the principle of such a diagram. It shows how the hazard loss of control – also referred to as a critical event – can be prevented or mitigated by appropriate countermeasures. For this, it focuses on the risk lead by vegetation growing in the neighborhood of power lines, lists potential threats and consequences, and enumerates existing barriers.[1]

Figure 1: Simplified Illustration of a Bow-tie Diagram



Multiple standards (e.g. *CSA Q850-97*, *ISO 31000:2009*, *NORSOK Z-013*)[2] support the process of risk assessment by providing frameworks and guidelines depicting the best practices in this field. These represent a solid basis for companies that aim to reduce the impact of unwanted and uncontrolled factors on their business.

Depending on what is at stake, power industry risk is expressed in different and often partial ways. For instance, potential socio-economical costs for stakeholders in the case of adverse event provide inputs for decision makers [14]. However, they only address the consequence factor from a financial perspective. Similarly, other methodologies focus on field-oriented indicators by evaluating the physical consequences of an event on the different assets present in the grid [2].

[1] The number of threats, barriers and consequences has been limited for consistency.
[2] An extended list is available at www.ntnu.edu/ross/info/standards

Different types of management systems (e.g., energy management systems, asset management systems, outage management systems) have been set up to minimize the occurrence and consequences of negative outcomes, as well as to ensure an optimal demand/supply-storage balance. These tools are related and affect each other's performance. Data management systems reflect the infrastructure life-cycle and are the core information sources to consider in risk monitoring [3]. They allow creating proxies to assess the performance of such countermeasures.

3.2. Shift to Smart Grids

Guidelines and practices underscore the importance of business knowledge, identification of environmental factors affecting the integrity of facilities, context knowledge of occurred events, information communication, continuous update, and, most importantly, risk reduction. In that sense, Smart Grids represent a great opportunity. Improved information transmission and automatized recovery processes aim to reduce frequency and duration of outages.

Smart Grids imply potential integration of new barriers as well as proxies. Nevertheless, in the current situation, grid companies face compatibility problems and monitoring challenges. Especially, it is difficult to integrate different systems, processes and datasets.

Moreover, Smart Grids may represent further vulnerabilities, which lead to the risk of new unwanted scenarios, such as hacking – key-names in the history of energy-related hacking are Stuxnet (Worm), BlackEnergy (Trojan), Industroyer (malware framework), DragonFly (Hacking group). As the privacy of customers represent a main concern due to the increase of hacking probability, a lot of attention has been given to cyber-security and cyber-attacks when it comes to risk management of Smart Grids.

3.3. The Common Information Model (CIM)

The *Common Information Model* (CIM) addresses the topic of intersystem communication. It was developed by EPRI in the 90's and aims to enable exchanges of information between the different systems involved in a power grid. It is an object-oriented standard – based on *Unified Modelling Language* (UML) – which aims to represent the different objects of interest, as well as their relations in terms of electric generation, distribution, transmission [15]. The development of this model is ensured by the *IEC Technical Committee 57*, especially through the development of the standards *IEC 61970: Common Information Model (CIM)/Energy Management*, and *IEC 61968: Common Information Model (CIM)/Distribution Management* which are also closely related to the standard *IEC 61850: Power Utility Automation*, focusing on the communication networks and systems in substations and power utility systems. The book of *Uslar et al.* "The Common Information Model CIM" [6] as well as the website of IEC[1] and the one of the CIM User Group[2] provide an interesting basis for anyone looking for additional documentation on the topic.

The broadening of CIM among stakeholders in the power industry has been facilitated in 2009, as the *European Networks Transmission Operators – Electricity (ENTSO-E)* approved CIM as a standard exchange format. As a broadly accepted model in the field of energy management, the CIM became an important element for the digital connection between datasets. However, acquiring the knowledge enabling to obtain information out of the data lake created remains a challenging task. Moreover, modelling, integration of information, model implementation and result interpretation still require important research progress to be effective.

[1] www.iec.ch/smartgrid/standards
[2] http://cimug.ucaiug.org

3.4. Related Work

Most of the methods currently used for risk assessment are based on the N-1 criterion [1], [2], [14], [16]–[19]. To paraphrase the sources referenced, this means that the companies in charge of the power transmission or distribution have to be able to ensure full power delivery, even if a main element of their network came to fail. The main downside of such deterministic methods is the lack of flexibility. Fine variations of uncertainty may be disregarded, while binary behavior may be represented by considering all events as equally likely.

Condition based risk assessment and other probabilistic methods represent interesting alternatives. As these methods base their efficiency on the amount of input data, they have logically suffered from poor computing performances of past IT technologies. Recent ICT evolutions led to new relevant projects [2].

Several EU projects (e.g. *AFTER*, *iTesla*, *Umbrella*, *HyRiM*) have focused on the topic of risk in electrical grids. *GARPUR*, one of the most recently completed project (November 2017), focused on asset management policies and outage scheduling assessment. Table 2 lists the main recommendations and calls for new research reported in the different deliverables of the project[1].

Table 2: Main Recommendations/Calls for Research of the *GARPUR* Project

N°	Recommendations/Calls for Research
1	Collect more asset-related and outage-related data (**failure rates**, degradation rates, estimation of expected life time, outage data, context dependent outage data, **costs, consequences**, impact of pre/post-event maintenance tasks, **restoration rates, failures of corrective controls/actions**) to improve currently applied models.
2	Add new types of data if the ratio (complexity added/information gained) is interesting; evaluate how these can be clustered with existing datasets.
3	Diversify imagined scenarios; better anticipate future challenging situation; consider large scale as well as more located scenarios (relevant in the context of micro-grids).
4	Increase the knowledge on smart technologies newly integrated into the grid; analyze the new features they enable; examine the way they interact with other devices; evaluate the way they should be modelled.
5	Consider environment-affected variable failure rates to better model the effects of slowly evolving conditions (e.g. pollution, dust, corrosion, vegetation).
6	Improve the quality and the availability of the data to gather.
7	Improve the coordination between Transmission System Operators (TSOs); Improve the coordination between Distribution System Operators (DSOs); Improve the coordination between TSOs and DSOs.
8	Sustain the definition of standards; improve the knowledge-sharing of standards.
9	Determine the value of potentially unreliable data (in the context of reliability metrics) by applying sensitivity analyses.
10	Define frameworks and guidelines for collecting, maintaining, and sharing data and models
11	Deepen the research for the automatization of processes related to risks by using machine learning techniques.
12	Sustain a gradual development, implementation and testing of methods relevant to the definition of a framework, rather than to wait for a final holistic solution to be proposed.
13	Develop new tools and sustain the formation of cross-disciplinary experts capable of providing risk-oriented analyses of heterogeneous datasets.

One of the initial and core topics to address is the access to data. As it is also further stated in the literature [10], [16], [17], [20] there is need for research on what type of datasets to gather, as well as a need for clarification on the way to intelligently combine data sources for the improvement of existing models in the field of risk assessment. Nevertheless, no clear guidelines have been found on the way to merge a large number of heterogeneous sources of information for improved risk assessment in the context of power grids - and more especially in the context of Smart Grids.

[1] www.sintef.no/projectweb/garpur/deliverables

The lack of maturity of methods focusing on degradation processes over long time periods and focusing on impact evaluation (positive-negative) of maintenance activities on components life-time highlight furthermore the need for more research in the field of probabilistic methods [1]. In order to provide such knowledge, research needs to focus on the nature of the relations existing between various phenomena and components' behavior to better consider obvious links (e.g. effects of storms on power lines) as well as slighter links (e.g., effects of dust, rust, and pollution on insulating properties) [10] in risk assessment and monitoring.

4. INVESTIGATION PROCEDURE

Considering the situation described and the challenges faced, we aim, with this work, to reduce the gap existing regarding risk assessment in Smart Grids. For this reason, we identify and report a list of data sources that can be used to better characterize risks. Moreover, we intend to highlight the links existing between the different sources in order to understand how the data should later be aggregated.

A deepened review of storm and outage reports, as well as an intensive research among the existing literature and among online websites of power management stakeholders has allowed identifying main elements and factors involved in the emergence of risks, outages and accidents in power grids. This research has enabled to identify main categories of directly related data sources, as well as tools that are used to reduce the severity of such outcomes.

An additional analysis has ultimately consisted in identifying the data sources that could have increased the knowledge level in one of the negative situations encountered, by providing identical, improved or complementary information, compared to the data initially considered.

5. RESULTS

Data and services initially defined for a specific purpose may lead to additional interesting features [10]. For instance, correlations among datasets enable creation of proxies and finer situation understanding. Table 3 provides a non-exhaustive list of data sources that could be used for more effective risk assessment. It illustrates the benefits of such information when combined with asset management and daily monitoring data from the grid, such as:

- grid topology & asset information (age, location, failure rates, initial life-time models, etc.),
- inspection and maintenance reports (preventive/corrective approaches),
- outage reports (context dependent outage data, consequences, costs),
- customer feedback and crew management decisions,
- power-flow forecasts & real-time power flow measurements in the grid,
- past, present and future design documents of the power grid.

The two first columns of table 3 identify the field of information considered; the third column highlights a relevant use case and proposes an access to data; and the last column suggests additional datasets to use for supporting the use case highlighted.

Table 3: Useful Sources of Information for Improved Risk Assessment

N°	Topic	Use Case	Add. Dataset
A	Vegetation	Anticipate risk induced by trees along power lines using growth models, health analysis and clearcutting reports. **Data access:** https://kilden.nibio.no	B, C, E, F, I, J, Q
B	Meteorological data, lightnings	**Historical data:** Identify risky areas (wind, ice-rain, etc.) for the planification of power grids - **Real-time data/forecasts:** Team management & material supply before and after events (snow, storms, etc.). **Data access:** www.yr.no, https://api.met.no, www.senorge.no	A, C-F, M, N, P
C	Satellite images	Enable real-time & post-event situation analysis; identify potential location of outages. **Data access:** https://open.nasa.gov, https://scihub.copernicus.eu	A, B, D, F, M, N
D	Region dependent disaster data	Evaluate risk level due to landslides, avalanches, floods, earthquakes, hurricanes, etc., for the planification of power grids. **Data access:** https://atlas.nve.no	A-C, E, F
E	Climate change forecasts	Anticipate snow melting for dams' reserves; Anticipate pests migrations. **Data access:** https://climate.copernicus.eu/seasonal-forecasts	B, C, F, I, M, N, Q
F	Terrain model, mineral composition	Anticipate evolution of streams and random water flows during floods; prevent short-circuits due to salinity. **Data access:** https://hoydedata.no/	A-E, J, N, P, Q
G	Forest fires	**Historical data:** Identify risky areas for the planification of power grids - **Real-time data/forecasts:** Team management & material supply during and after events. **Data access:** http://gwis.jrc.ec.europa.eu/	A-C, E, F, H, I, M, N, Q
H	Pollution, dust, smoke	Evaluate short-circuit probabilities and degradation due to aggressive chemicals [18], [21]. **Data access:** www.environment.no	B, C, M, N, P, Q, S
I	Animals, insects & fungi	Predict species migrations to estimate and respond to risk induced by animals coming in contact with electrical components (e.g. birds, squirrels) [22] or by pests deteriorating the infrastructures. **Data access:** https://gd.eppo.int/	A-C, E, H, M, N, P, Q, S
J	Laws & standards	Anticipate security requirements. **Data access:** List of standards, national/international regulations	B, D, E, H, K, L, O, Q-S
K	Union trade contest	Anticipate strikes. **Data access:** Internal data	J, L, M, O, R, S
L	Security and surveillance	Protection against potential espionage/sabotage/terrorism [16], [21]. **Data access:** Internal data, external audit reports, national security reports	C, K, M-R
M	Social networks	Anticipate risks due to social contest (organization of protest-events), detect/locate outages with clusters of dissatisfaction-hashtags, manage emergency-related outages (live messages, pictures and videos). **Data access:** Public feed API (Facebook), PowerTrack API (Twitter)	J-L, N, Q-S
N	Collaborative platforms	Increase risk-related knowledge thanks to voluntary contribution of anonym public. **Data access:** www.regobs.no	B-D, M, Q
O	Market, finance	Anticipate partners/customers behaviour due to market price fluctuations. **Data access:** Internal/partner/customer data	B, G, J-M, R, S
P	Simulated environment	Improve risk models thanks to simulated environments and simulated data. **Data access:** Internal/partner/customer data	B, D, F, Q
Q	Open-access data sources	Improve company's knowledge thanks to open-data politics of governments and external companies. **Data access:** https://open-power-system-data.org, https://rte-opendata.opendatasoft.com, https://transparency.entsoe.eu, www.data.gov	/
R	Geo-political information	Anticipate variations of power costs and consumption due to geo-political conflicts (e.g., Europe and its dependency to Russian gas). **Data access:** Various newspapers & other media	J, L, M
S	Other energy related data	Anticipate variation of power costs and consumption due to outages/problems in other energy utilities (e.g., explosion on a gas pipeline during the winter which leads to a pic of power consumption [21]). **Data access:** Internal/partner data, media	B-G, J-O, Q, R

6. DISCUSSION

The results highlight the plurality of data sources that can actually be relevant for risk assessment. Doing so, they enable to imagine the diversity of existing possibilities for the creation of proxies; which is a main asset for the increase of resilience in a power grid. The sample of suggested scenarios also highlights the plurality of the applications a data source can be useful for. Future work will need to focus on additional scenarios and need to look for more links across the datasets. This implies cross-disciplinary teams and the possibility to access and exchange datasets among the principal stakeholders.

It should be mentioned that the order in the table does not reflect the importance level of the data sources for the estimation of risk level, which is a matter of geographical localization and contextual configuration. Moreover, for the purpose of this work, a focus has been set on data relevant for a Norwegian power company. Some of the datasets suggested are thus not directly relevant for companies located outside of Norway and additional sources of information would have to be found in such circumstances. Finally, it is known to the authors that there is already room for optimization in the manipulation of the data used for asset management and daily monitoring. As, in such a case, the focus has to be set on the model and not on the access of data – which goes beyond the scope of this paper – this task has not been further deepened here.

Even if some online sources of information are known to be relatively less reliable (especially social networks and collaborative platforms) we stress that such sources should not be ignored for risk assessment, but rather integrated with precautionary methods enabling to estimate the value of the information before adding it into the models.

The manual combination of datasets described in this work is a pre-step for automated merging, which is relevant in the context of machine learning. However, further work still needs to be addressed before accessing to this automatization phase. A focus has especially to be set on aggregation models, which will enable to provide the first outputs directly utilizable by decision makers for the management of the power grids. This implies nevertheless to overcome many remaining challenges (variety of temporal resolutions/boundaries, variety of spatial resolutions/boundaries, restriction policies for data access, etc.) and calls for further research regarding data management solutions in Smart Grids.

7. CONCLUSION

Smart Grids have shown to be particularly promising. They come nevertheless with numerous challenges, from which we listed the most important ones. In this paper, we mentioned multiple relevant projects and publications focusing on risk assessment in modern power grids. We also brought forward a wide range of types of data sources that should be considered in this context and suggested different scenarios utilizing such sources of information to reduce the global level of risks in electrical grids. We finally highlighted which data sources could be combined to obtain more accurate estimations of actual risk levels existing in a real-world situation in the framework of Smart Grids.

With our work, we provide pieces/elements for the construction of a framework supporting better decision making in the management of risk in modern power grids. In order to provide actionable intelligence, there is nevertheless a need for better sharing of best practices and for better sharing of data. This will enable the principal stakeholders to get a more accurate overview of their infrastructures and so to better make decisions when it comes to risk management in their power grids.

Acknowledgements

This work is part of the project "*Dynamic risk management for Smart Grids in large-scale interconnected power systems.*" funded by *eSmart Systems* and the *Norwegian Research Council* (NæringsPhD program - Project No.: 276404), which the authors would like to thank for their support.

References

[1] M. Rekinger, M. Hofmann, A. Chaouachi, E. Karangelos, L. Wehenkel, and G. Kjølle, "A Transition Roadmap Towards Probabilistic Reliability Management," 2017.
[2] S. A. B. De Almeida, R. Pestana, F. P. M. Barbaosa, and N. Machado, "Operational Risk Assessment Methodology for the Portuguese Transmission System" *Electra*, no. 252, p. 11, 2010.
[3] N. Paltrinieri and F. Khan, *Dynamic Risk Analysis in the Chemical and Petroleum Industry: Evolution and Interaction with Parallel Disciplines in the Perspective of Industrial Application*. Elsevier, 2016.
[4] O. Nusbaumer, "Introduction to Probabilistic Safety Assessments (PSA)." p. 20, 2010.
[5] ETP SmartGrids, *European Technology Platform Smart Grids: Vision and Strategy for Europe's Electricity Networks of the Future*, vol. 19, no. 3. 2006.
[6] M. Uslar, M. Specht, S. Rohjans, J. Trefke, and J. M. Vasquez Gonzalez, *The Common Information Model CIM*, vol. 66. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
[7] R. W. Y. Habash, V. Groza, D. Krewski, and G. Paoli, "A Risk Assessment Framework for the Smart Grid," in *2013 IEEE Electrical Power & Energy Conference*, 2013, pp. 1–6.
[8] National Institute of Standards and Technology, "NIST Framework and Roadmap for Smart Grid Interoperability Standards, Release 3.0," 2014.
[9] S. Borlase, *Smart Grids: Advanced Technologies and Solutions*, Second. CRC Press, 2017.
[10] S. R. Khuntia, J. L. Rueda, and M. A. M. M. Van Der Meijden, "Smart Asset Management for Electric Utilities: Big Data and Future," in *WCEAM 2017*, 2017.
[11] L. Al-Hakim, *Challenges of Managing Information Quality in Service Organizations*. Idea Group, 2007.
[12] S. Kaplan and B. J. Garrick, "On The Quantitative Definition of Risk," *Risk Anal.*, vol. 1, no. 1, pp. 11–27, 1981.
[13] C. Delvosalle, C. Fievez, A. Pipart, and B. Debray, "ARAMIS project: A comprehensive methodology for the identification of reference accident scenarios in process industries," *J. Hazard. Mater.*, vol. 130, no. 3 SPEC. ISS., pp. 200–219, 2006.
[14] L. Wehenkel, M. Hofmann, A. Vafeas, E. Karangelos, and E. Jordanger, "Results and recommendations towards stakeholders," 2017.
[15] E. C. for E. S. (CENELEC), *Energy Management System Application Program Interface (EMS-API) - Part 301: Common Information Model (CIM) Base (IEC 61970-301:2011)*. 2011, p. 253.
[16] S. Perkin *et al.*, "Modelling weather dependence in online reliability assessment of power systems," *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.*, vol. 231, no. 4, pp. 364–372, 2017.
[17] V. V. Vadlamudi, C. Hamon, O. Gjerde, G. Kjolle, and S. Perkin, "On Improving Data and Models on Corrective Control Failures for Use in Probabilistic Reliability Management," in *2016 International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2016 - Proceedings*, 2016.
[18] N. Machado, S. A. De Graaff, and R. Pestana, "Risk assessment methodology – Running tests at the Portuguese TSO," in *CIGRE 2014*, 2014.
[19] O. Gjerde, A. Vafeas, C. Badajoz, S. Dourlens-Quaranta, and S. H. Jakobsen, "Final Conference Proceedings," 2017.
[20] TECHNOFI, "Technical analysis of on-going projects," 2015.
[21] NERC, "State of Reliability 2017." p. 208, 2017.
[22] NRECA, "Animal-Caused Outages," 2005.

Article III - Supporting the Application of Dynamic Risk Analysis to Real-World Situations using Systems Engineering : A focus on the Norwegian Power Grid Management

Pacevicius, M., Haskins, C., Paltrinieri, N., 2020. **Supporting the Application of Dynamic Risk Analysis to Real-World Situations using Systems Engineering : A focus on the Norwegian Power Grid Management**, in: 18th Annual Conference on Systems Engineering Research - Recent Trends and Advances in Model-Based Systems Engineering. Virtual, p. 8.

18th Annual Conference on Systems Engineering Research
Recent Trends and Advances in Model-Based Systems Engineering

Eds.: Azad M. Madni, Barry Boehm
Daniel A. Erwin, Mahta Moghaddam, Michael Sievers, and Marilee J. Wheaton
University of Southern California
October 08 - 10, 2020

Supporting the Application of Dynamic Risk Analysis to Real-World Situations using Systems Engineering: A focus on the Norwegian Power Grid Management

Michael Pacevicius[a,b,*], Cecilia Haskins[b], Nicola Paltrinieri[b]

[a]*eSmart Systems, Håkon Melbergs vei 16, 1783 Halden, Norway*
[b]*Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, Trondheim, Norway*

Abstract

Dynamic Risk Analysis (DRA) approaches are virtuous processes enabling the improvement of state-of-the-art techniques for risk calculation in industrial infrastructures. However, they require the existence of an appropriate architecture enabling end-to-end processing of information, which has not yet been defined in practice. This paper aims at discussing the possibilities and the advantages of combining DRA with Systems Engineering (SE) approaches to reach this objective. For that, we define a framework based on SE principles, apply it for the assessment of the role of vegetation on the global risk for power grids and discuss the benefits it provides.

Keywords: Dynamic Risk Analysis; Systems Engineering; Internet-of-things; Power Grids; Vegetation; Critical Infrastructures

1. Introduction

System managers aim, among other objectives, to reduce uncertainties related to process monitoring and to maximize the control efficiency. Those dimensions are, for the most part, defined by the capacity of the stakeholders involved to properly analyze, evaluate and reduce the risk levels characterizing the system under review. Numerous guidelines and standards have been developed to support these activities. However, the way

[*] Corresponding author. Tel.: +47 919 16 430
E-mail address: michael.pacevicius@esmartsystems.com

such guidelines are used and the way they affect the quality of monitoring activities show a considerable heterogeneity across fields of application. In addition, the nature and the specificities of the supporting tools also influence the performance level of risk analyses.

Although most risk assessment methods advocate for feedback loops and cyclic processes, the quality of the outputs of these procedures tends to deteriorate over time. This is especially due to the inability for currently implemented tools to integrate in a continuous, reliable, and thus dynamic, way information related to the system under review. Researchers aiming to tackle this dilemma have suggested frameworks and approaches leading to more cyclically effective, and thus dynamic, risk analyses and evaluations (Villa et al., 2016). The application of such approaches is still in an embryonic phase in real world scenarios. This is mainly due to the relative recency of the proposed tools, which still require recommendations for applicability. On the other hand, although such techniques advocate for a better and more intense use of information, they do not describe technical solutions able to properly integrate all the data that is theoretically available and that is suggested to be analyzed in the recommended approaches. In fact, the applicability of such approaches is highly dependent on the existence of an architecture to support this integration in a timely manner. This is especially challenging, considering the general complexity of the systems under review and thus the plurality of influencing factors and data sources to consider.

Systems Engineering (SE) in general offers interesting perspectives for addressing this challenge. By fully considering local specificities while keeping a global understanding of the needs for which a system is developed and the constraints it is and will be subject to, SE offers development perspectives enabling designers to properly attack the resolution of the problem. The present paper discusses the advantages systems engineering can provide to support the application of Dynamic Risk Analysis (DRA). These advantages are demonstrated by an application in risk analysis done for power grids.

2. Risk Analysis: Original Concepts and Requirements for Dynamic Evolutions

Risk is a variable concept across domains due to the different ways people and industries value situations and the consequences of events. The way risk is defined and integrated into models/tools depends on the type of business (Aven, 2012), which has led multiple standards to be developed within the diverse fields of application where risk understanding and control is required (ISO - International standardization organization, 2007, 2016, 2018; NORSOK, 2010).

Generally, the first step for risk analyses consists in properly defining the context for the entire life cycle of the system under review. This includes qualitatively and quantitatively defining the measures of effectiveness relative to risk acceptance, as well as the way performance measures will be verified/validated all along and at the end of the process. This step aims to eventually obtain a risk picture that accurately depicts the true exposure level of the system to specified threats. The risk picture is characterized by the contribution of all scenarios leading to an unwanted event or situation, the likelihood for these scenarios to occur and the severity of the negative consequences resulting from these scenarios (Kaplan and Garrick, 1981).

The tools used for risk analysis in industrial facilities have been observed to have the following limitations:

- Recurrent use of outdated data for frequency evaluation, despite the acquisition of new knowledge based on experiences (Creedy, 2011),
- No capture of interactions and dynamic aspect of risk variations (Yang and Haugen, 2015),
- Inappropriate consideration of uncertainties related to risk (Aven, 2012; Villa et al., 2016).

Furthermore, despite the theoretical inclusion of a cyclic feature in most risk assessment approaches, experience shows that the proper reassessment of establishing the context suffers from some latency as the number of cycles increases. This initial step is often done once for all at the beginning of the study; which means that further steps of the analyses are not able to react correctly as internal or environmental conditions of the system evolve over time. In addition, accurately characterizing the evolution of the hazards that can impact the system becomes challenging. Thus, although the awareness of requirements for flexibility and adaptability is generally present in the conception phase, the implementations usually suffer from an inability to update and integrate new information. Consequently, the estimated level of risk may, over time, diverge from the true level of risk, which in turns may lead to new potential accidents or catastrophes.

The appropriate processing of correct information – and more especially the capability to consider new variables or be resilient to disturbances (e.g., loss of data source) – represents an area that needs more research. Acknowledging this situation, more dynamic methods have been developed (Villa et al., 2016). In general, these methods require:

- real-time acknowledgment of information updates for initially considered variables,
- the capacity to integrate (or discard) variables and thus restructure the method when new (or old) information is considered relevant (or irrelevant) for the estimation of the current risk image.

Thus, dynamic methods require the consideration and integration of data-driven updates rather than only considering long-term calendar-based protocols, which is a common practice across industries today. This means that optimal data management is at the heart of dynamic risk analysis. However, in order to make existing tools structurally updatable and adaptable, there is a need to offer a standardized approach that technically enables automatic integration or suppression of relevant data, thereby enabling evolution from isolated data sources towards informed risk depiction.

3. Key Dimensions of System Engineering and Contributing Potential to DRA

Properly controlling and monitoring a system over time requires an appropriate understanding of the composition of the system and the existence of an appropriate user interface. Furthermore, the behavior of the system needs to be understandable from both a local and a global perspective, in order to identify and assess interactions between sub-systems and their respective characterizing variables; thus allowing systems engineers not only to understand the sum but also the product of the sub-systems' respective behaviors. Pooling a team of specialists from a diversity of areas supports such actions, as it provides the analysis of multiple dimensions of the system and breaks barriers between fields presenting synergies.

SE advocates for cyclically reviewing and adapting, if necessary, the different phases of the process in order to show evidences that the system fulfils its functions as expected by the customers and stakeholders. Periodically carrying out context evaluation ensures a good understanding, over time, of the system's properties and interactions between its subsystems and helps to avoid the probability of misunderstood or wrongly quantified hazard effects.

The understanding of interrelations between variables and the cyclic requirement in the development of a system represent key dimensions for both efficient SE approaches and DRA tools. Table 1 details further the numerous correspondences which exist between DRA requirements and SE specificities.

Table 1. Correspondences between DRA requirements and SE specificities.

Requirements for efficient DRA tools	Specificities of Systems Engineering
(1) The structure and the architecture of the tools enabling optimal data management in complex environments need to be defined.	(1) SE approaches support the construction of complex systems in an efficient and durable way.
(2) DRA tools need to be updatable in terms of architecture.	(2) Proper system design enables flexibility in the architecture of the system and ability to consider new variables as required.
(3) DRA tools need to be resilient to degradation or loss of data sources (quality & quantity of data).	(3) Proper system design enables to make efficient use of correlations between variables and thus provides "as good as knowable" analyses.

Requirements for efficient DRA tools	Specificities of Systems Engineering
(4) Cyclic assessments and reconsiderations of context are required to maintain process understanding and thus to keep a realistic risk picture over time.	(4) Integration of a cyclic dimension is a pillar of SE which reduces the non-detections of emerging hazards or the apparition of black boxes within the process.
(5) Interactions between factors need to be captured to show accurate risk pictures.	(5) Implementation of interdisciplinary approaches is a pillar of SE.
(6) Frequently updated data need to be considered to show accurate risk pictures.	(6) The cyclic dimension of SE supports the frequent updating of information.
(7) Use of case-/plant-specific data.	(7) Correctly designed systems integrating appropriate data sources enable "as good as knowable" analyses.

4. Approach Description

The approach suggested in the present paper is based on the steps of regular data mining flow processes (Chapman et al., 2000) and is structured using a Systems Engineering mindset in order to provide an efficient analysis of risk over time. It is developed as following:

(1) Identification of information requirements (Business Understanding):

Within the context, the different needs of the system stakeholders are gathered. This enables understanding which type of information needs to be available to provide an appropriate solution, as well as to direct the first steps of the research.

(2) Identification of potentially accessible data sources (Data Understanding):

Based on the requirements formulated by the stakeholders, establish a benchmark of the existing data sources (or of data sources that can be created to reach the defined objectives). The type of data sources corresponds to data sources that, somehow, by their nature, enable a better understanding of the analyzed items. Only those datasets that are accessible for the project are retained.

(3) Filtering of data sources (Data Understanding, Data preparation):

The informative potential of the retained information sources is initially assessed and used to create a maximum number of scenarios to consider in terms of risk. Discussions between the heterogeneous panel of experts involved enables understanding the importance of the physics for each observed variable, but also – and maybe more importantly – the interrelations and dependencies that can exist between considered variables/phenomena. Data sources providing quantitatively usable information (e.g., databases with numerical values) or convertible (e.g., by some weighting conversion process), are then selected. The informative potential of each of the related retained variables is then assessed in terms of contribution to the calculation of the targeted risk dimension. More especially, each of the variables is evaluated to assess if they provide information relative to the frequency or the consequences of the identified scenarios. Based on the importance of the identified scenarios, requirements for the data acquisition of the related variables is estimated and reported for future performance evaluation of the defined system.

(4) Clustering of data sources (Data preparation):

The data sources that are considered are then clustered, based on their resolution and on the reported physical interrelations existing between the observed variables.

(5) Choice of potential environments, frameworks, and algorithms (Modelling):

The choice of the environment to work in, as well as the frameworks and algorithms is based on both the objectives to achieve and the characteristics of the retained data sources.

(6) Structuring of the pipeline (Modelling)

The clusters suggested in phase (4) are integrated into layers in which analyses will be done. Layers with lowest resolution are placed in an initial position of the pipeline for optimization of the workload management. The selected frameworks and algorithms are adapted for each layer, depending on the fixed objectives.

(7) Progressive use of outputs and assessment of pipeline final results (Evaluation, Deployment)

Once the pipeline is prepared, each layer is successively run through in order to eventually reach the final risk picture. Estimations originating from the output of the successive layers is also sent via feedback loops in order to improve the algorithms exploited via approval or rejection of the first results.

A critical requirement of the approach suggested is a good understanding of the techniques used. This will be the only way to properly convert the information they provide into usable inputs for the improved calculation of risk levels.

5. Application: Pipeline Construction for Improved Risk Analyses in Power Grid Management – Focus on Vegetation

5.1. Situation overview

The power grids used daily are exposed to a plurality of hazards (e.g., hurricanes, earthquakes, ice storms, floods), which occurrences can have heavy consequences (DeCorla-Souza, 2013; Kenward and Raja, 2014). In addition, dimensions such as the size of the grid, the accidental terrain it can be installed in and the slow, local and complicated processes used to gather information for inspections and maintenance mean that exercises related to risk analyses often are executed in a sub-optimal way. By suggesting a Systems Engineering-based approach, we aim to show how the general level of risk in power grids can be reduced in a continuous way, giving thus evidence that more dynamic approaches can be implemented. For this purpose, we focus here on the impact of vegetation on the power grid. Vegetation was the number one cause of outage in Norway in 2018 (Eggum, 2019) and is a main contributing factor for outages in power grids in general (Hansen, 2018, 2017, 2016). The most common way for vegetation to affect the power grid is generally by a branch or an entire tree falling directly on a power line. Alternatively, vegetation can also generate outages by simply growing under a line until it makes contact and creates an outage. In the best-case scenario, consequences of such events can be relatively low, with only a few power customers affected. However, such events can also lead to wildfires (Kumagai et al., 2004) or contribute to large blackouts (Alhelou et al., 2019).

Multiple parameters are involved in the occurrence of an outage generated by a tree falling on a power line. The first, obvious ones, are the size of the trees and their physical proximity to the power line. Additional factors are wind or precipitations, variations in temperature, the topography, the species, health and shape of the trees - to name just a few.

Grid operators require the following information for decision-making concerning vegetation management:

- Locations of areas that are more likely to face outages involving vegetation to send teams clear-cutting the region before there is a problem;
- Level of consequences of such an outage when it happens (particularly in terms of impacted customers);
- Location of areas that are more likely to face outages involving vegetation to know where to look first when those occur, and thereby shorten reaction time and eventual power restoration.

5.2. Proposed architecture

The architecture proposed to tackle the problem of vegetation is divided into two main phases, as described in figure 1.

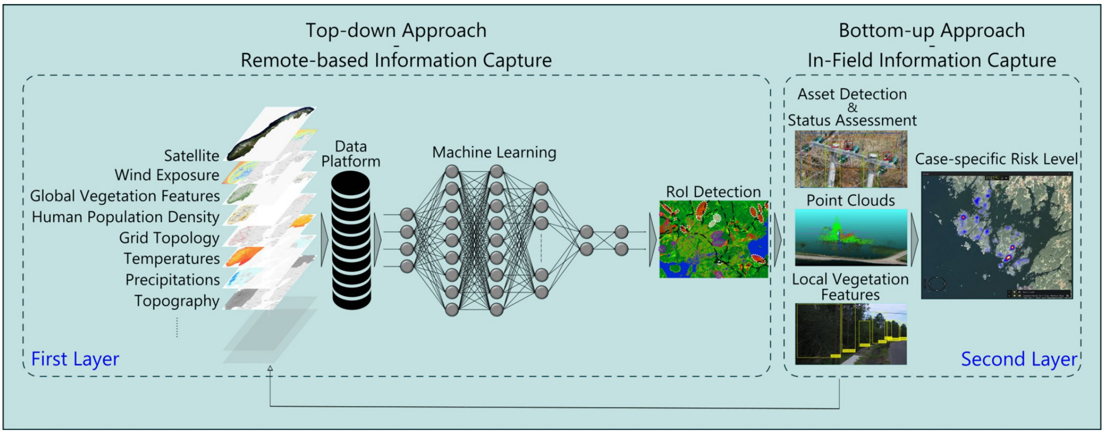


Fig. 1. Architecture for vegetation-focused Dynamic Risk Analyses in Norwegian power grids

In a first phase, the first layer, essentially a top-down approach, is initially a "remote-based" information capture. It integrates diverse sources of information, such as optical satellite images, wind exposure, global vegetation characteristics [i.e., dominant species presence, canopy height], human population density, grid topology, temperatures, precipitations, and topography. Information relative to each data source is first collected and stored in a database in such a way that it can be used for calculations. Machine-learning based methods enable the combination of the different variables to assess and report the contributing factor of each variable into the first layer risk calculation. This first layer risk calculation aims to provide, by combining both the probability and the consequences of a potential disturbance, estimations of regions where the levels of vegetation-based risk estimation is the highest (Regions of Interest – RoI).

Although calculations are case-specific, results of risk estimation remain characterized by a medium degree of uncertainty because of the resolution of the information used. In order to accept or reject the first level estimation and thus reduce the level of uncertainty around the risk estimations, a second layer (which represents a bottom-up approach) is introduced in a second phase. This "in-field-based" information capture is here again the fruit of several fields of expertise and enables specific assessment of assets and detection of faulty components using computer vision; centimeter-level distance estimation from trees to power grid components using lidar-based or photogrammetry-based point clouds; and tree specific characteristics estimation [i.e., species, height, health conditions] using computer vision. Computer-vision based asset information suggests if the probability of outage might be increased by the types and condition of the asset present in the specified RoI. Point clouds enable case-specific high-resolution distance measurements from vegetation to the power line. And computer-vision based tree characterization enables assessment of how the originally estimated level of risk may be affected by the properties of the trees present in the area. These three additional data sources contribute by providing case-specific local information, offering an improvement of both the consequence estimations and the probability estimations of an outage in the RoI, thus enabling a refinement of the final risk image.

The output of the second layer is feedback to the first layer as the re-assessment cycles unfold and helps improve the quality of some first estimations of the local area (e.g. trees height and species), which enables the inclusion of high-resolution time-series into the calculations and improved estimations provided by the retrained algorithms.

5.3. Results

The described architecture enables the operator to:

- Discover previously unconsidered risky areas and thus better quantify the consequences of disturbances caused by potential outages;
- Multiply the number of scenarios leading to an unwanted event by highlighting relevant interactions between relevant variables;
- Refine existing scenarios and risk contribution levels of specific variables by enabling a higher resolution situation understanding;
- Refine estimation of the contribution level of each factor to the global risk picture;
- Increase resilience to loss of information by increasing the number of data sources. This increases the probability of correlation detections among used data sources, which can thus be used as proxies when one/some of them fail or would be removed;
- Make case-specific risk estimations/improvements by gathering local data, avoiding use of averaged values and reducing thus uncertainty around risk estimation;
- Increase the frequency of risk estimations by benefiting from regular updates (e.g. weather) of the data sources used for the risk estimation.

Based on this dynamic risk estimation, power grid companies can improve the assignments of woodcutting teams by efficiently prioritizing missions based on potential risk, thereby reducing the occurrence of vegetation-influenced outages. Furthermore, they can reduce the time to repair if an outage happens since a product of this architecture increases the probability of spotting the correct areas causing the outage.

The key dimension of the described architecture is to make “as-good-as-knowable” estimations, optimizing the contributing potential of the accessible data sources, increasing the probability of detection of early signals and reducing the probability of occurrence of events that can be avoided with timely use of information.

6. Discussion and Conclusion

Although the demand and justifications for the development of DRA tools is obvious across industries, it remains a challenging task and a relatively new research area. A theoretical broad access to a large number of data sources and an easy access to powerful IT infrastructures suggest that the main entities that could support the development of DRA tools are already available in practices that support the emerging Internet-of-Things. Structuring the combination of those different entities and transforming this combination into a useful risk image for an infrastructure under review remains a challenging task requiring competences in a multitude of disciplines, a local understanding of the interrelations, as well as a holistic overview of the constructed system, considered within a specific environment. SE is a particularly supportive field with this regard, as it provides the right framework to develop white-box-based systems for which understanding and control can be kept over time.

We illustrated the benefits that can be provided by SE for the development of DRA tools by focusing on the assessment of the impact of vegetation on the true risk level existing in power grids. For that, we showed how relevant data sources should be combined in such a way that decision makers can optimize their judgements and the management of their resources, as well implement a pre-event resilience plan, and effective post-event restorations. How the risk reduction actions are executed and the way the resulting information is integrated into the new cycle of risk assessment is an additional dimension that needs to be carefully considered to ensure optimal risk management of the infrastructure.

The expected convergence between requirements for efficient DRA tools and solutions provided by the intrinsic properties of Systems Engineering is confirmed in the proposed approach and described in the case-study. By enabling the creation of systems that favor cyclic approaches, SE enables a flexible process and offers possibilities for both optimized information management and more resilience. This makes the process more reliable, sustainable and thus suitable in the long run for the application of Dynamic Risk Analysis.

This work is part of an ongoing project "*Dynamic risk management for Smart Grids in large-scale interconnected power systems*" funded by eSmart Systems and the Norwegian Research Council. Future steps include the final choice of the best algorithms for the machine-learning processes capturing those variation rates of the related data sources. Those elements will be the basis for complementary performance metrics (cf. phase (3) of the suggested approach) and will enable quantification of the level of uncertainty of the calculated risk level.

The generic dimension of the proposed approach offers already a flexibility that enables it to be used for other industries, under the condition that informative data sources are properly identified and accessible. The approach also requires the validation and verification to be continuously ensured by close collaboration with the customer, a condition that has been respected in the present study. Continuously exchanges with the main stakeholders enables appropriate feedback with regards to the performance, the process design and the context evolution. In this way, the best trade-off options can be continuously chosen and divergence between the suggested risk representation and the real risk level over time can be reduced.

References

Villa, V., Paltrinieri, N., Khan, F., Cozzani, V., 2016. Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. Safety Science, 89, 77–93.

Aven, T., 2012. The risk concept - historical and recent development trends. Reliability Engineering and System Safety, 99, 33–44.

ISO - International standardization organization. ISO 14971:2007 - Medical devices - Application of risk management to medical devices, 2007. Geneva, Switzerland.

ISO - International standardization organization. ISO 17666:2016 - Space systems - Risk management, 2016. Geneva, Switzerland.

ISO - International standardization organization. ISO 31000:2018 - Risk management: principles and guidelines, 2018. Geneva, Switzerland.

NORSOK. Standard Z-013 - Risk and emergency preparedness assessment, 2010. Lysaker, Norway.

Kaplan, S., Garrick, B. J., 1981. On The Quantitative Definition of Risk. Risk Analysis, 1(1), 11–27.

Creedy, G. D., 2011. Quantitative risk assessment: How realistic are those frequency assumptions? Journal of Loss Prevention in the Process Industries, 24(3), 203–207.

Yang, X., Haugen, S., 2015. Classification of risk to support decision-making in hazardous processes. Safety Science, 80, 115–126.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0 - Step-by-step data mining guide. P. 76.

DeCorla-Souza, K., 2013. Comparing the Impacts of Northeast Hurricanes on Energy Infrastructure, p. 50.

Kenward, A., Raja, U., 2014. Blackout: Extreme Weather, Climate Change and Power Outages. Climate Central, p. 23.

Eggum, E., 2019. Rapport Nr. 29-2019 - Avbrotsstatistikk 2018. Oslo, p. 101.

Hansen, H., 2018. Rapport Nr. 64-2018 - Avbrotsstatistikk 2017. Oslo, p. 103.

Hansen, H., 2017. Rapport Nr. 43-2017 - Avbrotsstatistikk 2016. Oslo, p. 108.

Hansen, H., 2016. Rapport Nr. 78-2016 - Avbrotsstatistikk 2015. Oslo, p. 99.

Kumagai, Y., Bliss, J. C., Daniels, S. E., Carroll, M. S., 2004. Research on causal attribution of wildfire: An exploratory multiple-methods approach. Society and Natural Resources, 17(2), 113–127.

Alhelou, H. H., Hamedani-golshan, M. E., Njenda, T. C., Siano, P., 2019. A Survey on Power System Blackout and Cascading Events Research: Motivations and Challenges. Energies, 12, 1–28.

Article IV - Optimizing Technology-based Decision-support for Management of Infrastructures under Risk: The Case of Power Grids

Pacevicius, M., Ramos, M.A., Paltrinieri, N., 2020. **Optimizing Technology-based Decision-support for Management of Infrastructures under Risk: The Case of Power Grids**, in: 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference. Venice, p. 8.

OPTIMIZING TECHNOLOGY-BASED DECISION-SUPPORT FOR MANAGEMENT OF INFRASTRUCTURES UNDER RISK: THE CASE OF POWER GRIDS

Michael F. Pacevicius

eSmart Systems, Norway.
Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, Norway. E-mail: michael.pacevicius@esmartsystems.com

Marilia A. Ramos

The B. John Garrick Institute for the Risk Sciences, University of California Los Angeles UCLA, United States. E-mail: marilia.ramos@ucla.edu

Nicola Paltrinieri

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, Norway. E-mail: nicola.paltrinieri@ntnu.no

Technological developments allow for gathering and processing an increasing amount of data in real time. The integration of these tools into risk assessment allows for the development of dynamic risk assessment and data-driven decision support. The latter is of special interest for systems that are remotely monitored and controlled by operator, such as power grids. Generally, grid operators have little access to environmental information to support decisions on interventions and preventive maintenance. Recent initiatives aim at integrating machine learning and other techniques into dynamic risk assessment of power grids. The performances of these initiatives depend on the quantity and quality of data one can gather and process, the available technology, and the cost-benefit ratio of which these initiatives are synonym. In addition, the development of these solutions must be completed by the list of decisions to which the operators may be subject, as well as the information required in order to make the correct decisions for the system's needs. This paper presents a framework for optimizing decision-support of power grids operators using data-driven solutions, focusing on risks associated to vegetation. We analyse the possible scenarios concerning power grids under risk by surrounding vegetation, and the deriving decisions the operators can make under those scenarios. We further analyse and discuss the information required by the operators for decision making. This information is finally integrated into a data-collection and processing framework.

Keywords: Dynamic Risk Analysis, Decision Support, Power Grid Management, Vegetation Hazard, Preventive Maintenance

1. Introduction

The advent of machine learning, big data, internet of things and other technological solutions, enabled by the increasing access to powerful machines, allows for gathering and processing data in real time. The integration of these tools into risk assessment, in turn, allows for the development of dynamic risk assessment and data-driven decision support. The latter is of special interest for a diversity of applications. Indeed, systems that are remotely monitored and controlled by operators can highly benefit from technological solutions for increasing operators' situation awareness and keeping them in the loop. As such, data-driven and risk-based decision support can thus be particularly beneficial for management of power grids.

Risk management is a main concern for decision makers in power grid related companies. To provide end-users with a reliable and continuous energy flow, they need to ensure the functioning of the grid at all times. Yet, the power grids are exposed to a plurality of hazards such as hurricane, earthquakes, ice storms, floods, etc. Those hazards can have severe consequences (DeCorla-Souza, 2013; Kenward and Raja, 2014). In addition, the vulnerability to these hazards can increase depending on the terrain on which the power poles are installed, the remoteness of the power grid and its size, among other factors.

One of the main hazards related to power grids operations is vegetation. Vegetation can affect a power grid in case a branch or an entire tree falls directly on a power line, or in case it grows under a line, making contact and creating an outage. In some cases, the consequences of these events can be relatively low, with only a few power customers affected. However, consequences can also be particularly severe, such as wildfires (Kumagai et al., 2004) or large blackouts (Alhelou et al., 2019, Sforna and Delfanti, 2006). Vegetation was the number one cause of outage in Norway in 2018 (Eggum, 2019) and is a main

Proceedings of the 30th European Safety and Reliability Conference and
the 15th Probabilistic Safety Assessment and Management Conference.
Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio

ISBN: 981-973-0000-00-0 :: doi: 10.3850/981-973-0000-00-0 esrel2020psam15-paper

contributing factor for outages in power grids in general (Hansen, 2018, 2017, 2016).

Vegetation management is typically performed through growth assessments and clear-cutting operations. This can be a costly and challenging task. For instance, internal information reported by an European power grid company shows that a simple visual inspection made by an operator in a helicopter can already have an approximate cost of 40€ per kilometre. The locations and frequency of clear-cutting operations are, in addition, decided based on limited data. Yet, several data sources are available and can be leveraged for risk-informed decisions, such as satellite images, generalized use of point clouds, open-access vegetation-related databases, etc.

The use of these data sources in power grid management is suggested in (Pacevicius et al., 2018). In order to develop an effective decision-support, the data use and technology development can be shaped considering the type and format of the data that is required by the operators for risk-informed decisions.

This paper presents a framework for a risk-based decision-support for operators managing power grids under vegetation hazard. The decision-support is data-based, and benefits from current technological developments. The framework includes the assessment of vulnerable areas and the possible consequences in case of vegetation hazards. The combination of vulnerability and consequence levels results in a risk-ranking of areas that should be visited for clear-cutting. Further development and application of the framework can result in cost-effective and efficient preventive maintenance, reducing outages and other consequences.

This paper is structured as follows: Section 2 presents an overview on power grids operations, followed by a discussion on data driven solutions for power grids management in Section 3. Section 4 presents the framework for decision-support. Finally, concluding thoughts are presented in Section 5.

2. Power grids operation

Power grids are managed by either Transmission System Operators (TSO) or Distribution System Operators (DSO). These are the operational and regulatory bodies that share the responsibility of properly installing, managing and maintaining the power grids. Despite the long experience these operators have in the field, they continue to face major disturbances all over the world, with sometimes particularly damageable consequences (Alhelou et al., 2019). This is particularly true with exceptional meteorological events. For instance, hurricanes Irene (2011) and Sandy (2012) resulted in 6,69 million and 8,66 million people without power respectively (United States Department of Energy, 2013). More recently, over 6 million people lost power in 2017 due to hurricane Irma (NERC, 2018).

The challenges concerning power grid management are not restricted to abovementioned large impact hazards. Power grids' intrinsic characteristics also pose daily challenges. For instance, the network dimension implies a challenging configuration of large and complex systems leading to dynamics hardly forecastable. This favours the occurrence of unpredictable cascading events, which can have large consequences if occurring on a favourable terrain. In addition, the broad geographical distribution of the power grid mathematically increases the exposure of the infrastructures to external threats. This can make the surveillance, control and management of the grid particularly complicated, especially in remote areas.

Power grids' intrinsic features impose notable difficulties concerning execution of management related tasks, such as maintenance operations and resource optimization. For example, the size of the power grid directly impacts the way inspection and maintenance tasks are scheduled and executed. Those can be time consuming (some inspections tasks are planned on a 10-year basis calendar) and risky, especially in mountainous regions such as in Norway.

The importance of power grids and the nature of the threats to which they are exposed call thus for continuous application and improvement of proper risk management methods. State-of-the-art methods suggest the use of Dynamic Risk Analyses (DRA) for fulfilling this task, which enable a better exploitation of information over time for an improved understanding of the true risk level (Villa et al., 2016).

The potential benefits of using DRA for power grids management are strengthened given the increasing technology for collecting and processing environmental and meteorological data, as well as information concerning users, grids' physical conditions, and others. Next section provides an overview of the available data and its potential use, as well as the related challenges.

3. Overview of potential data driven solutions for power grid management

Data-driven dynamic risk management methods enable to consider changes in key variables as new information is made available. Static approaches, on the other hand, make use of pre-defined plans for scheduling inspection and maintenance tasks. DRA approaches are data driven methods relying on a capacity to consider both (1) real-time data updates for variables already at scope and (2) real-time variable updating, as previously identified hazards may change in terms of potential impact and/or new hazards may emerge. Variable updating implies both integrating new variables if those have been proven to contribute to the risk depiction and/or ignore variables which have become irrelevant for the analysis in process. Proper data management (selection, collection, process and update) is thus a cornerstone for the existence of DRA. The implementation of DRA methods is strongly empowered by the general proliferation of interconnected IT-based technologies, commonly referred to as the “internet of things”.

In the field of power grids, the broader use of interconnected devices and the capacity to access more data sources has led to the emergence and extension of “smart grid” configurations, principally over the two last decades. Furthermore, power grids have gained in connectivity since the implementation of the first SCADA (Supervisory Control And Data Acquisition) systems, and management of large-scale infrastructures has been improved over the years thanks to the integration of new technologies.

3.1 Possible data sources

The first cluster of data sources to consider for the evaluation of vegetation-related outages in the power grids is common to all type of outages and composed of the systems enabling to observe the power flow variations. Based on sudden interruptions in the power delivery, operators will be informed of the occurrence of an outage and start to investigate its causes, consequences and location. Depending on those results, they are able to look for an alternative solution in order to return as quickly as possible to initial service levels via rerouting of the power delivery. SCADA systems support those type of operations and are integrated into the management of power grids since multiple decades already. The granularity level of the data that is nowadays accessible has however strongly increased since the implementation of the first SCADA devices, moving from on overview on region level, towards the possibility of obtaining an understanding of local substation and finally the ability to observe power variations on individual building level thanks to the generalized implementation of Advanced Metering Systems (AMS), also called “Smart Meters”.

Operators also use different methods to acquire information regarding vegetation. The most common approaches consist in visual inspection such as foot patrols, helicopters, and drones (Nguyen et al, 2018). Maintenance reports summarize the main conclusions of these inspections. Additionally, Light Detection And Ranging (LiDAR)-based point clouds are frequently used to get precise distance measurement between power grid infrastructures and foreign objects such as vegetation. This technique is however relatively costly, limiting its use in practice.

A benchmark of the technologies and existing solutions for supporting vegetation-related risk analysis was explored by (Pacevicius et al., 2018). The most relevant results for this paper can be summarized as follows:

- Satellite images: the multiplication of image providers has considerably increased over the last years, leading to both a strong increase of performances (higher resolution images for different bands can be obtained more frequently) and a strong decrease in costs (down to a few dozens of euros/square kilometre, depending on technology and resolution).
- Weather: meteorological models have gained in accuracy over the last decades, enabling to make local estimations of the wind exposition, exposition to precipitations and temperature variations.
- Topography: Digital Terrain Models (DTM), more and more accessible, enable to obtain a description of the raw surface on which power grids are installed, enabling to estimate the role of the terrain orientation (i.e., slope) in the stability of trees.
- Referencing of species: some national registries report the mostly present species per area, enabling to better estimate the average stability of that part of the forest in the area.

Combining the different data sources mentioned, with the grid topology, the database of supplied customers and related levels of importance (e.g., individual housing, hospital, data centre, aluminium factory, etc.), there is here a strong potential to affine the estimation of both the probability and the consequences of a outages caused by vegetation.

3.2 Related challenges

In spite of recent improvements, the implementation of smart grid technologies and infrastructures still face many remaining challenges (Pacevicius et al., 2018). These can be illustrated by data gathering: several advances have been done on the hardware side, while merging and processing data from different data sources is still challenging. Indeed, many sensors have been developed and installed into the infrastructure, leading to the possibility to acquire a large quantity of data. Intra-disciplinary and highly specialized research studies have also enabled progresses in a plurality of fields (e.g., weather forecasts, physics of electricity, computer vision, satellite data analysis). Yet, those advances have mostly been done in silos. The added value of merging inter-disciplinary knowledge remains difficult to be acquired because of a lack of method and recommendations enabling these combinations. This is partially explained due to the challenges of processing heterogeneous data sets (Pacevicius et al., 2018) and due to the complexity resulting from the combination of fields of expertise.

An additional challenge for effective use of data, results from the methods used for its collection. For example, the processing of the information gathered during inspections remains often slow, local and complicated. This is mainly due to the use of outdated and paper-based methods in the

treatment of maintenance reports. In cases more advanced methods for data collection are used, such as using tablets for filling maintenance reports, a lack of proper procedure hinders the use of the data in an automatic and efficient way. Once saved, digital maintenance reports remain usually exploited by other operators in a manual way, similarly to paper-based reports. Transmission of the information may thus, in the best-case scenario, be facilitated from one device to another (e.g., one tablet to another), but its exploitation remains mostly manual and rarely efficiently automated. This illustrates the gap between the ever-increasing data availability and the techniques enabling to exploit them.
Making a proper use of the generated data in order to be aware of real-time of situation evolution (and hence, make risk informed decisions) is thus a challenging task. Automating transmission and processing of data to provide high quality information to decision makers requires:

- a good business/ scenario understanding in order to know which information is looked for,
- the implementation of hardware enabling the acquisition of good quality and relevant quantities of data,
- the existence of field-specific expertise to pre-process acquired data,
- the understanding of inter-disciplinary dependencies,
- the knowledge to merge resulting information, and
- the capacity to know which type of decision can be taken in the different identified scenarios.

Those elements correspond to the building blocks of a system architecture enabling the sustainable development of DRA methods. Recent study by (Pacevicius et al., 2020) stresses the need of the development of feedback loops methods for enabling the consideration and integration of new observations made by the systems and/or the operators. Above all, the architecture must be developed in such a way that efficient information use can be made possible by power-grid operators for improved decision making. It must thus be profitably integrated in their daily tasks, considering their expectations and the challenges they face.

4. Integrating data driven solutions for optimized power grids decision support

Power grids operators make decisions on preventive and corrective maintenance on a daily basis. Today, these decisions are based on limited information. This is especially true because of the latency sometimes resulting from slowly transmitted information in current day-to-day operations, where some asset-related databases are only manually updated after several years (Pacevicius et al., 2018).
When more detailed information is necessary, such as the state of vegetation close to a power line, in-loco data must be acquired. These operations can be highly costly and error prone.
The use of the data and technology described in the previous section can improve the efficiency of decisions on preventive maintenance. The level of vulnerability of a power line and the possible consequences in case of an outage generated by vegetation hazard can provide the operators with priority areas for observation and clear-cutting. Those, in turn, can highly improve resource allocation while avoiding outages and more severe consequences.
This section presents a framework for data-driven decision-support for power grids operators. The scope of the framework is on vegetation hazard and preventive maintenance. The framework can be expanded for other hazards types. The problem statement is described in Section 4.1, followed by an analysis of the operators' decisions in Section 4.2. These are integrated with data-driven technologies in Section 4.3 for the decision-support framework.

4.1 Scenario description: vegetation hazard

Outages due to vegetation are mostly due to branches/trees falling on the infrastructures or due to vegetation growing under/on/in the installations, generally leading to short circuits and/or damaging of infrastructures. Vegetation-related outages represent a cost to the grid operator, in terms on non-delivered power to its different customers. Financially, it also implies having to send teams on the ground to clean the affected area and replace the parts of the power lines that have been damaged during the event. Sending teams solving such situations in not without significance and does sometimes lead to tragical consequences (Line, 2016). In addition, disturbances may lead to reputation loss. Above all, vegetation hazard can lead to wildfires. In case the location and environmental conditions are favourable, these can quickly spread and have severe outcomes.
Several variables can indicate the vulnerability of a power line to vegetation hazard. Some are related to the trees, such as their size, species, health and shape, and their proximity to the power line. Additionally, external parameters such as wind exposition, exposition to precipitations, temperature variations, and topography, are known to affect the stability of trees and should as such be studied for a proper vegetation-related risk analysis.

4.2 Power grids' operation and vegetation hazard

Power grid operators can be sectorized in different groups, depending on their role in the management of the network. Although operators' roles and responsibilities can be changing among different organizations, a common distinction can be described as following:

(1) Control-room operators: those operators oversee real-time management of the power grid and focus on a short-term horizon. They are responsible for restoring the initial service levels as fast as possible after an outage has occurred, among other tasks.

(2) Maintenance operators: They are responsible for the management of inspections & maintenance operations in order to avoid the occurrence of outages and potentially reduce their consequences if those would occur. As such, they work on a short- to middle-term horizon.

(3) Planning operators: They are responsible for future extensions or rerouting of the power lines when new customers integrate the network or when alternatives for existing routes are required.

Although the distinction above may not be the same among all organizations, this paper adopts it for increasing clarity and comprehension in this case study.

Current operations relative to vegetation management are characterized by two main limiting factors:

- Maintenance operator schedule inspections and preventive vegetation cleaning operations on a calendar basis rather than on real needs;
- Control-room operators are aware of an incident only after it has happened.

A risk-based decision support can give grid operators the possibility to optimize the currently applied decision-making. With an indication of vulnerable areas and their risk levels, operators can prioritize crucial operations and send teams for clear-cutting the area where required before any damage is reported, postponing in parallel non-urgent inspection missions. In addition, they will be able to spot areas more likely to be affected by vegetation-related outages when disturbances are observed, increasing thus the probability to gain precious minutes in the power grid restauration.

The use of data and related technology for decision-support is generally performed as in Figure 1. The available data-driven solutions are assessed based on the possible information they can generate. This information, in turn, is assessed based on its capacity to support operators' possible decisions. Yet, the development of the decision-support can follow a different path, as we suggest here and as is illustrated in Figure 2. The operators' possible decisions are initially defined in view of the system's needs (i.e., vegetation management in the present use-case). The information required for the possible decisions is then further identified. The necessary information for decision making is finally used to shape the development and exploitation of data-driven solutions. Note that the present process is a high-level approach brought forward to illustrate the need of shaping data driven solutions by firstly considering operators' needs. Development of data-driven solutions are however also known to be shaped by hardware and software limitations, as well as by the possibilities to access relevant databases.

Figure 1: Use of data and related technology for decision-support

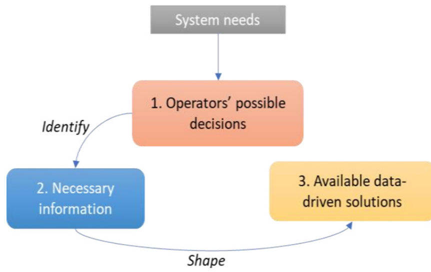


Figure 2: Approach for development of data-driven decision support shaped by operators' needs

The first step into developing an effective decision support is thus to evaluate the operators' possible decisions when managing vegetation hazard. Three possible decisions are considered:

(1) Send a team for observation

In case the information provided by the system has a high level of uncertainty, operators can send observation teams to assess vegetation level and reduce the uncertainty. The system should be designed to accommodate the output of this operation and use it to update the level of vulnerability of the grid.

(2) Send a team to clear-cut the area

The operators should receive enough information to decide when it is necessary to send a team to clear-cut the area, i.e., perform preventive maintenance. This decision can be made considering the risk level provided by the system.

(3) Redirect flow/shut down grid sections

This decision is to be made in extreme situations, in which the operator believes required maintenance may not be done in time and the potential consequences resulting from the reported top events may be too severe.

A next step is to evaluate the necessary information for the operators to make the correct decisions. The decision of sending a team for observation operations is based on the reported exposition level of the grid to the vegetation-related hazards, as well as on the uncertainty degree of this estimation.

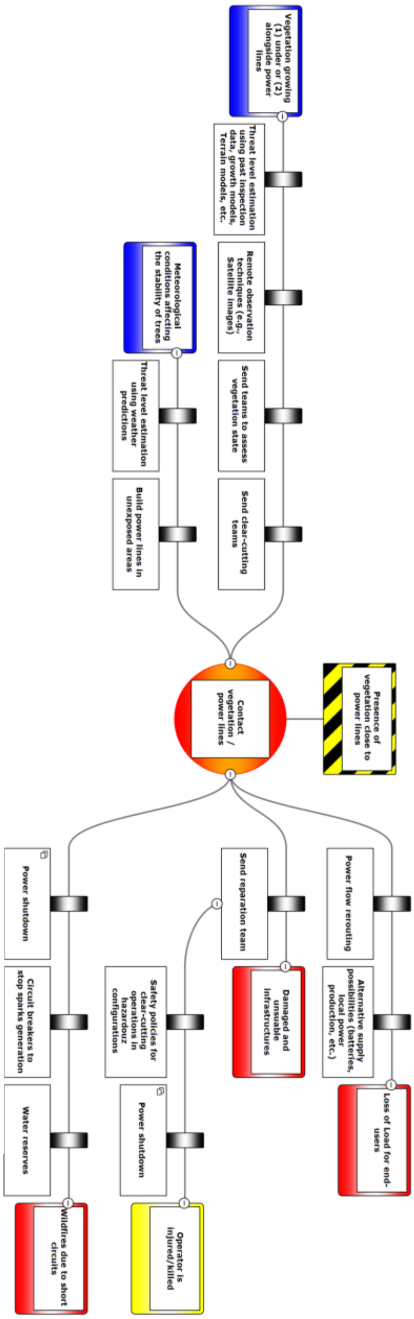


Figure 3: Bowtie diagram describing the most relevant causes & consequences of vegetation coming in contact with power lines

A decision to sending a team for clear-cutting the area can moreover be taken when consequences of potential outages are furthermore added to the present analysis. This is particularly important for resources allocation and actions prioritizing, especially in case two or more vulnerable spots are reported. Finally, the more extreme decision of redirecting power flows and/or shutting down power, relies on the previously mentioned sources of information, plus the overview on the available and missing resources in terms of personnel, material, power budget, time, etc.

A partial bowtie diagram addressing the main causes & consequences relative to the vegetation-related hazard is available in figure 3. It is based on (Pacevicius et al., 2018) and focuses on the most relevant dimensions in the present study. It is however to be mentioned that additional elements (i.e., barriers, threats and consequences) may be added, depending on the case study. The present diagram - as well as the data sources used to exploit it - can furthermore be used to shape the development of technology and data-driven solutions, as further explored in the following section.

4.3 Decision support framework

Risk can be defined in multiple ways, depending on the fields to which it is applied, the methods it is integrated in or the authors it is used by (Aven, 2012). We adopt one of the most common definition, provided by Kaplan and Garrick (Kaplan and Garrick, 1981), in which *risk* is as a triplet of a *scenario* (s) happening, as well as the *probability* (P) and the *consequences* (C) of this scenario occurring. Furthermore, due to its relevance to the present work, we add the variable of *uncertainty*. The risk level to be used by the operators for decision support is thus a function of:

$$R=f(s,P,C,u)$$

In the scenario of vegetation hazard, the probability refers to the probability that the vegetation under or around the power line affects the grid. The probability will indicate the vulnerability level of the line. This probability is a function of the vegetation specificities, its proximity to the power lines, the environmental conditions and the local shape of the terrain:

$$P=g(\text{vegetation specificities, distances to the lines, environmental conditions, terrain})$$

The consequence of the event refers to the possible outcomes in case vegetation affects the line, as described in figure 3. The consequences are estimated based on the impact the realization of the different scenarios have on the involved assets. By converting big/important costumers (hospitals, large industries, etc.) into an equivalent "number of individual households" in terms of importance (e.g. 1 hospital ≈ 500 individual households), one can quantify the impact relative to non-delivered power in terms of "non-supplied end-users". Focusing on wildfire, we can integrate a binary variable characterizing each Are in the direct surrounding of the lines as being prone to wildfire (1) or not (0). Finally, one should also integrate infrastructure costs related to corrective maintenance. In the present case, the consequences can thus be expressed as:

C=h(∑ [non-supplied end-users], ∑ [Are prone to wildfires], infrastructure costs)

The uncertainty, in turn, refers to the uncertainty associated to the generated probabilities and consequences.

The way the different data sources can be combined to obtain the mentioned risk metrics and eventually support power grid operators in the decision they take is illustrated in Figure 4. A relevant use-case based on ongoing work[1] can illustrate the framework as following:

<u>Probability of outage:</u>

- o Using Satellite images and point clouds, one can automatically detect the presence, size and proximity of trees in the surrounding of power lines.
- o Adding topographical data and vegetation categorization, one can assess the importance of relatively static influencing factors.
- o Completing the analyses with weather-related and power flow-related time-series, one can obtain a dynamic estimation of the probability of an outage happening.

<u>Consequence of Outage:</u>

- o Equivalent number of non-supplied end-users,
- o Quantity of energy not provided,
- o Values of infrastructures,
- o Number of Ares prone to wildfires,

<u>Uncertainty:</u>

- o Based on the uncertainty related to the data sources and the process chosen to combine their output, one can eventually estimate the uncertainty level related to both the probability of outage estimation and the consequence of outage estimation

5. Discussion and Concluding Thoughts

The present paper illustrates the added value of integrating data-driven solutions into decision-support for management of infrastructures under risk. For that it takes the example of power grid management and focuses more especially on the evaluation of the threat originating from the presence of vegetation close to power grid infrastructures. It illustrates how relevant data sources should be manipulated in such a way that decision makers can optimize their judgements and the management of their resources, as well in a pre-event resilience perspective as from a post-event restauration point of view. How the risk reduction actions are executed and the way the resulting information is integrated into the new cycle of risk assessment is however an additional

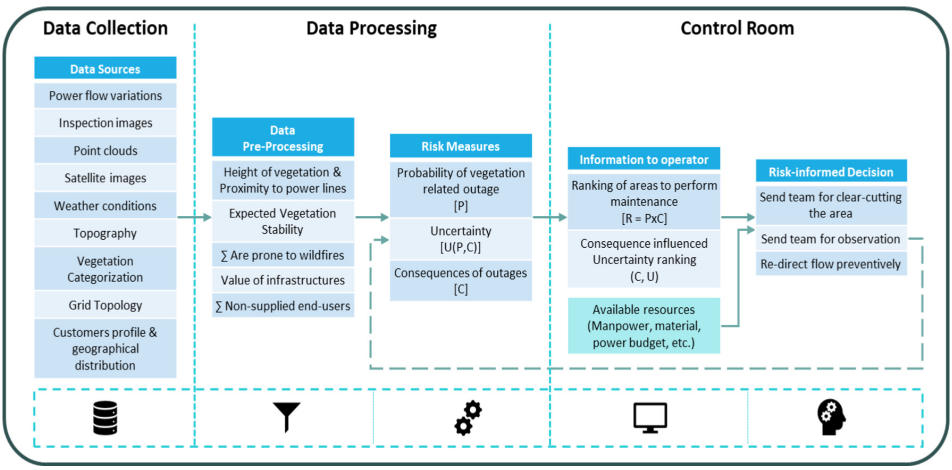


Figure 4: Framework for data-driven decision support for power grid operators for management of vegetation hazards

[1] https://business.esa.int/projects/grideyes

dimension that needs to be correctly executed in order to ensure optimal risk management of the infrastructure over time. While continuing the development of the presented framework and reinforcing the steps already suggested, there is thus a need for further studies in order to assess and quantify the role of human factors on the final true risk level, and evaluate how to reinforce the feedbacks of the operators in the developed system.

A data-driven and efficient decision-support as suggested by the framework can lead to correct decisions by the operators regarding maintenance scheduling and prioritizing. Yet, operators' decision is also influenced by internal factors (e.g. distraction, fatigue) and external and organizational factors (e.g. human-system interface quality, procedures, workplace adequacy). Those must also be considered when developing a system that aims to reduce human error.

With the description done in the present work, we focused on the risk generated by vegetation in power grids. Although vegetation is often the final element in the causal chain, the causes of events including vegetation are often plural and include, for example, wind or snow. Those factors taken independently can also be at the origin of outages, making them similarly items to assess in order to quantify their influence on the final holistic risk image. As such, further work will also need to deepen and highlight the interactions between variables in order to properly quantity the influence of each dimension on the true risk value.

This paper highlights the benefits of interconnected IT-based technologies for the management of infrastructures under risk, supporting that way the generalization of data-driven methods for risk analyses and calling thus also for a generalization of the use of dynamic risk analyses. The extended access to a large number of data sources can however also come with additional complications and lead to potential sources of inefficiency. As such, this paper is also an occasion to recall the need for risk assessment methods to be both model-based and data-driven.

Acknowledgement

This work is part of the project "*Dynamic risk management for Smart Grids in large-scale interconnected power systems.*" funded by eSmart Systems and the Norwegian Research Council (NæringsPhD program - Project No.: 276404), which the authors would like to thank for their support.

References

Alhelou, H.H., Hamedani-golshan, M.E., Njenda, T.C., Siano, P., 2019. A Survey on Power System Blackout and Cascading Events Research: Motivations and Challenges. Energies 12, 1–28. https://doi.org/10.3390/en12040682

Aven, T., 2012. The risk concept—historical and recent development trends. Reliab. Eng. Syst. Saf. 99, 33–44. https://doi.org/10.1016/j.ress.2011.11.006

DeCorla-Souza, K., 2013. Comparing the Impacts of Northeast Hurricanes on Energy Infrastructure.

Eggum, E., 2019. Rapport Nr. 29-2019 - Avbrotsstatistikk 2018. Oslo.

Hansen, H., 2018. Rapport Nr. 64-2018 - Avbrotsstatistikk 2017. Oslo.

Hansen, H., 2017. Rapport Nr. 43-2017 - Avbrotsstatistikk 2016. Oslo.

Hansen, H., 2016. Rapport Nr. 78-2016 - Avbrotsstatistikk 2015. Oslo.

Kaplan, S., Garrick, B.J., 1981. On The Quantitative Definition of Risk. Risk Anal. 1, 11–27. https://doi.org/10.1111/j.1539-6924.1981.tb01350.x

Kenward, A., Raja, U., 2014. Blackout: Extreme Weather, Climate Change and Power Outages. Clim. Cent. 23.

Kumagai, Y., Bliss, J.C., Daniels, S.E., Carroll, M.S., 2004. Research on causal attribution of wildfire: An exploratory multiple-methods approach. Soc. Nat. Resour. 17, 113–127. https://doi.org/10.1080/08941920490261249

Line, H.P., 2016. Forest Crew Worker Electrocuted While Trying to Cut Tree Fallen on.

NERC, 2018. Hurricane Irma Event Analysis Report 1–33.

Nguyen, N., Van Jenssen, R., Roverso, D., 2018. Automatic Autonomous Vision-based Power Line Inspection : A Review of Current Status and the Potential Role of Deep Learning. International Journal of Electrical Power & Energy Systems.

Pacevicius, M., Haskins, C., Paltrinieri, N., 2020. Supporting the Application of Dynamic Risk Analysis to Real-World Situations using Systems Engineering : A focus on the Norwegian Power Grid Management.

Pacevicius, M., Roverso, D., Rossi, P.S., Paltrinieri, N., 2018. Smart grids: Challenges of processing heterogeneous data for risk assessment. PSAM 2018 - Probabilistic Saf. Assess. Manag.

Sforna, M., Delfanti, M., 2006. Overview of the events and causes of the 2003 Italian blackout. 2006 IEEE PES Power Systems Conference and Exposition, 301-308. https://doi.org/10.1109/PSCE.2006.296323

United States Department of Energy, O. of E.D. and E.R., 2013. Comparing the Impacts of Northeast Hurricanes on Energy Infrastructure 50.

Villa, V., Paltrinieri, N., Khan, F., Cozzani, V., 2016. Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. Saf. Sci. 89, 77–93. https://doi.org/10.1016/j.ssci.2016.06.002

Article V - Lessons from past hazardous events: data analytics for severity prediction

Paltrinieri, N., Patriarca, R., Pacevicius, M., Salvo Rossi, P., 2020. **Lessons from past hazardous events: data analytics for severity prediction**, in: 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference. Venice, p. 8.

Lessons from past hazardous events: data analytics for severity prediction

Nicola Paltrinieri

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, S. P. Andersens veg 3, Trondheim, Norway. E-mail: nicola.paltrinieri@ntnu.no

Riccardo Patriarca

Department of Mechanical and Aerospace Engineering, Sapienza University of Rome, Via Eudossiana, 18, Rome, Italy. E-mail: riccardo.patriarca@uniroma1.it

Michael Pacevicius

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, S. P. Andersens veg 3, Trondheim, Norway. E-mail: michael.f.pacevicius@ntnu.no

Pierluigi Salvo Rossi

Department of Electronic Systems, Norwegian University of Science and Technology NTNU, O. S. Bragstads plass 2, Trondheim, Norway. E-mail: pierluigi.salvorossi@ntnu.no

An increasing amount of information is collected by the monitoring systems within the process industry, especially concerning safety management. For instance, the Seveso III regulation on the control of major-accident hazards involving dangerous substances is the first version that refers to the collection of safety indicators for monitoring the performance of safety management systems. This leads to a call for improvement in learning past lessons and definition of techniques to process relevant data, in order to deal with unexpected events and provide the right support to safety management. Through this work, we suggest a data analytics approach for severity prediction of future hazardous events. The approach is twofold and is based on the use and comparison of multiple linear regression (MLR) and deep neural network (DNN) models. These models are developed and tested on the Major Hazardous Incident Data Service (MHIDAS) database. A set of simulations has been carried out not only to evaluate the models, but also to identify their limitations. The results show the capability of these models to manage heterogeneous data from past accident records and extract important information to support safety-related decision making. It must also be mentioned that intrinsic model limitations should be considered, and appropriate model selection and customization should be carefully carried out to deliver the needed support.

Keywords: learning from lessons, data analytics, multiple linear regression, deep neural network, safety management.

1. Introduction

Hazardous events may manifest under various forms in industry, but they mostly known as events involving the loss of containment of hazardous materials in the process industry (Pasman, 2015). In Europe, the handling of hazardous materials by process industry is regulated by the so-called the Seveso III regulation on the control of major-accident hazards involving dangerous substances (European Parliament and Council, 2012). As stated by Pasman, the loss of control of such substances has the potential to cause high-impact low-probability accidents (Pasman, 2015). High impact indicates catastrophic losses, but, due to their low probability, these accidents may even not happen during a lifetime.

Paltrinieri et al. (Paltrinieri et al., 2013, 2011) address another aspect of major accidents, as, in some cases, they are the results of scenarios that are "not captured by hazard identification methodologies because deviating from normal expectations of unwanted events or worst case reference scenarios." These may occur when hazard identification does not produce a complete overview of system hazards (Paltrinieri et al., 2010).

Another term used to define rare catastrophic events that have never been encountered before was coined by Taleb (Taleb, 2007), who used the metaphor of the Black Swans. These events can be explained only in the aftermath and cannot be anticipated, such as the black swan was believed to be impossible before its discovery in the 17th century (Taleb, 2007). However, the concept may be misused as it may represent a reason for

Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference
Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio
Published by Research Publishing, Singapore.
ISBN/DOI: TBA

ignoring the potential for major accidents and avoiding the implementation of long-term safety measures (Paté-Cornell, 2012).

The concept of Dragon-Kings (Sornette, 2009), may indicate a responsible approach to deal with major accidents. Dragon-Kings are defined as events that are extreme and outliers (in analogy with the kings' wealth), but unlike anything else, such as dragons. These major accidents are intended as the result of some degrees of organization and coordination of relatively smaller unwanted events and features, which could serve to amplify the final consequences.

Extreme accidents are the result of a combination of such details, some of which may be considered as deviations from normal/optimal conditions.

These deviations can be defined as early warnings (Paltrinieri et al., 2015a) or associated with the concept of "Small Things" (Paltrinieri and Khan, 2016). Small things might be recurring old issues in a plant or organization, which do not need imaginative definitions to be prevented, but perhaps only the compliance with already present procedures. Acting on Small Things would allow breaking the chain of events and lower the probability for major unpredictable accidents.

In the last decade, increasing attention has been dedicated to evaluation and monitoring of early deviations through appropriate indicators, to assess and control risk. Indicators can be represented by a series of factors: physical conditions of a plant (equipment pressure and temperature); number failures of an equipment piece; maintenance backlog; number of emergency preparedness exercises; amount of overtime worked; etc. Several indicator typologies have been theorized and used, but we often address risk indicators if (Øien, 2001): they provide numerical values (such as a number or a ratio); they are updated at regular intervals; they only cover some selected determinants of overall risk, in order to have a manageable set of them.

The latter feature has quickly become outdated due to the extensive collection that is being carried out in industry and the attempts made to process and elaborate larger numbers of them. For instance, for the first time since the first Seveso directive was issued in 1982, Seveso III mentions specific procedures for safety performance indicators and/or other relevant indicators, to use for monitoring the performance of safety management systems [3].

Table 1 reports how such suggestion has been received in some of the EU member and associated countries. Past hazardous events are collected by all the countries considered in Table 1. United Kingdom and France use specific databases to collect them. On the other hand, the use of safety performance indicators is not as common across Europe, but where it is not present, it is suggested by relevant research institutes.

Table 1. Seveso III-based monitoring approaches in the EU member and associated countries (Paltrinieri and Reniers, 2017).

	Indicators	
	Past events	*Safety performance indicators*
United Kingdom	Hazardous events reported to the competent authorities and regulated by RIDDOR (Reporting of Injuries, Diseases and Dangerous Occurrences).	The British competent authorities require hazard establishments to collect safety performance indicators (PSPIs).
France	Hazardous events are collected in the database ARIA (Analysis, Research and Information on Accidents). These events are also used as Key Performance Indicators.	The French national competence centre for industrial safety and environmental protection (INERIS) suggests the use of a Safety Performance Indicator System.
Italy	Hazardous events are reported to the competent authority.	The regulation states that safety performance monitoring should be based on indicators.
Netherlands	Hazardous events are reported to and collected by the competent authority.	The regulation requires companies handling hazardous substances to collect safety performance indicators.
Finland	The competent authority uses hazardous events to assess the performance of safety management systems.	Other indicators are also used to assess the performance of safety management systems.
Norway	The competent authority collects hazardous events according to the Seveso regulations.	The Norwegian research institute SINTEF suggests monitoring the safety trend of Seveso establishments with safety performance indicators.

This leads to a call for improvement in learning past lessons and definition of techniques to process relevant data, in order to deal with unexpected events and provide the right support to safety management. However, industrial risk analysis is unevenly progressing within this topic (Paltrinieri et al., 2019). At the same time, the use of machine learning has possibly become more attractive, given the progressive refinement of its models and the exponential increase in available computing power (Goodfellow et al., 2016). For this reason, we suggest a data analytics approach to predict the severity of potential hazardous events based on lessons learned from past hazardous events.

2. Method

The approach is twofold and is based on the use and comparison of multiple linear regression (MLR) and deep neural network (DNN) models. MLR and DNN may be considered as techniques belonging to the field of machine learning, which refers to techniques aiming to program computers to learn from experience (Samuel, 1959). While MLR is a statistical technique that uses several explanatory variables to predict the outcome of a response variable (Andrews, 1974), DNN aims to simulate (to a certain extent) the learning model of the human brain (Goodfellow et al., 2016). It is loosely based on information processing and communication patterns in a neural system. It allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.

A computer may be trained to assess risk or some of its components for safety-critical industries such as Seveso-regulated sites through machine learning techniques. This would allow processing the large amount of information currently collected in the form of lessons from past events or safety performance indicators. Moreover, although risk level cannot be evaluated with certainty, machine learning can allow for expert supervision through supervised learning (Goodfellow et al., 2016).

2.1 Multiple linear regression

A linear model, given a vector of inputs $X = (x_1, x_2, \ldots, x_p)$, predicts the output y (in this case an index for the risk R) via the following equation (Hastie et al., 2009):

$$y = b_0 + \sum_{j=1}^{p} x_j w_j \sim R \quad (1)$$

where b_0 is the so-called bias and wj represents the model weights. This model needs then to be trained with a training set data in order to learn the weights of every provided input. Once the weights are known, the model can be used for prediction of y based on new inputs X.

2.2 Deep neural network

The deep learning model considered in this work is a feed-forward neural network, wherein connections between the units do not form a cycle (Svozil et al., 1997). A linear model, such as MLR, would be restricted to linear functions, while a DNN model describes the target as a nonlinear function of the input features (Goodfellow et al., 2016). The DNN model can be described as a series of functional transformations associated to the model layers (Figure 1).

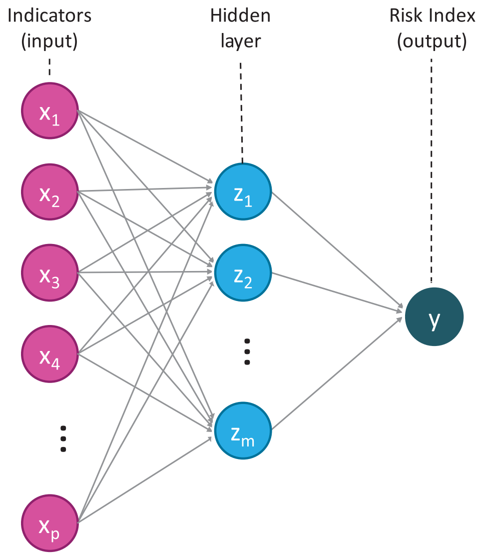


Fig. 1. DNN layers.

The overall length of the chain gives the depth of the model. Specifically, the first network layer performs the following computation of the inputs $X = (x_1, x_2, \ldots, x_p)$:

$$a_i = b_i + \sum_{j=1}^{p} x_j w_{i,j} \quad (2)$$

with $i=1, \ldots, m$.

Where a_i, b_i and w_i are respectively defined as activation, bias and model weight.
The activations are transformed by the activation function g within the hidden layer:

$$z_i = g(a_i) \quad (3)$$

Where z_i is defined as hidden unit. The most used activation function is the sigmoid (Goodfellow et al., 2016). Figure 1 shows only one hidden layer for the sake of simplicity, but there can be several.

The hidden units are combined to give the activations a_o of the output layer:

$$a_o = b_o + \sum_{j=1}^{m} z_j w_{o,j} \quad (4)$$

Where a_o, b_o and w_o are activation, bias and model weight. Figure 1 shows only one output for the sake of simplicity, but there can be several.

Finally, the activation function h is used to obtain the output y, which is an index for the risk R:

$$y = h(a_o) \sim R \quad (6)$$

Given a dataset of X and associated y, the model can be trained to minimize the final loss function in a supervised way, in order to predict y based on new inputs X.

3. Application

The described approach was applied to a database of past accidents with the purpose to simulate its application on the national databases managed by the Seveso-competent authorities. The dataset used is the Major hazard incident database (MHIDAS) (AEA technology - Major hazards assessment unit, 2003) launched by the UK Health and Safety Executive in 1986 and developed by AEA Technology until mid 1990's. The events included are based on public domain information sources and their characteristics are registered using keywords.

MHIDAS collects about 8972 hazardous events from 1916 to 1992, recorded by means of the set of items listed in Table 2. Some items use a taxonomy to systematically categorize the event.

Table 2. Set of items used to record hazardous events in MHIDAS (AEA technology - Major hazards assessment unit, 2003). Specific keywords are used to describe some of the items.

Items	**Description**	**Category from taxonomy**
Date	Date of the event	
Location	Location of event	
Substance	Substances involved in the event	X
Event type	Typology of event	X
Origin	Area of the plant and type of equipment from which the event started	X
Section	Plant section in which the event occurred	X
Quantity	Amount (ton) of released substance	
General causes	General causes the led to the event	X
Specific causes	Specific causes the led to the event	X
Evacuated	Number of people evacuated	
	Consequences	
Damage	Economic damage to the property or production loss	
Injured	Number of people injured by the event	
Killed	Number of people killed by the event	

The items listed in the upper part of Table 2 where considered as inputs X to the models, in order to predict the consequences – lower part of Table 2. The details of data pre-processing are explained elsewhere (Solini, 2017). The study focused on the number of people killed and aimed to predict the occurrence of a hazardous event within one of the severity categories listed in Table 3 based on the considered inputs. Only categorical data are used.

Table 3. Severity categories considered by the study.

Severity categories	
0	Event with no fatalities
1-10	Event with a number of fatalities between 1 and 10
10-100	Event with a number of fatalities between 10 and 100

Two datasets were created from the overall MHIDAS database:

1. A training dataset used to train the MLR and the DNN models, with 2/3 of the x_i and associated y values, and
2. A test dataset used to test the models, with about 1/3 of the x_i and associated y values.

A code in Python language was written for training and testing. The classifiers tf.contrib.learn.LinearClassifier and tf.contrib.learn.DNNClassifier from the open-source library TensorFlow (Google LLC, 2018) were used for the models. The DNN model structure (i.e. number of layers and nodes) was inspired by Cheng et al. (2016), based on which the hyper-parameters are defined.

4. Results

The results show whether the events from the test dataset were predicted within the correct severity category from Table 3. The models produce a probability of belonging to a severity category. The probability threshold based on which the decision on whether an event belong to a category is set at 0.5 by default. This affects the following cases:

- true positive (t_p,), as correct prediction of event belonging to a severity category;
- false positive (f_p), as incorrect prediction of event belonging to a severity category;
- true negative (t_n), as correct prediction of event not belonging to a severity category; and
- false negative (f_n), as incorrect prediction of event not belonging to a severity category.

In order to obtain an overall evaluation of the MLR and DNN prediction capabilities, the specific metrics listed in Table 4 were considered. While accuracy, precision and recall are defined based on the default threshold value, the area under the precision/recall curve (PR AUC) is calculated variating the threshold value from 0 to 1.

The results obtained generally show good capability to predict a hazardous event without fatalities in both the methods, as all the metrics reach values at about 0.8 or more.

Table 4. Metrics describing the prediction capabilities of the models.

	Definition
Accuracy	$Acc = \frac{t_p+t_n}{t_p+t_n+f_p+f_n}$
Precision	$Pr = \frac{t_p}{t_p+f_p}$
Recall	$Re = \frac{t_p}{t_p+f_n}$
PR AUC	The area under the Precision/Recall curve.

However, the prediction capabilities sensibly decrease in case of prediction of hazardous events within the severity categories involving fatalities. The only metric increasing for these categories is the accuracy, which almost reaches the unitary value for the category "10 to 100 fatalities". Precision and recall show values under 0.2 and next to 0 respectively for the categories "from 1 to 10 fatalities" and "from 10 to 100 fatalities". PR AUC maintains slightly higher values in both the categories involving fatalities.

5. Discussion

The results show the capability of two machine learning models to manage heterogeneous data from past accident records and extract important information to support safety-related decision making. In fact, the records of hazardous events reported on MHIDAS include both items described by a set of specific keywords and numerical values.

These data were used to build two parallel models predicting the severity of potential new hazardous events. The metrics obtained from testing the two models show good capabilities in predicting hazardous events without fatalities. However, it must be pointed out the presence of class imbalance as the events without fatalities represent the vast majority within the dataset considered. This could be the reason of such a variation in performance when we look at the remaining severity categories.

Hazardous events that cause from 1 to 10 fatalities were predicted with an accuracy that is over 0.8, while the category "from 10 to 100" show even higher accuracies. However, it must be considered that the accuracy metric represents a partial evaluation of the model. In fact, if the model is employed for the prediction of rare events (such as the ones with fatalities), predicting always their "non-occurrence" would lead to high accuracy anyway as the metric presents the term of "true negatives" at the numerator. This is demonstrated by the other metrics, in particular precision and recall, which are equal to 0 in the last category reflecting the absence of predicted "true positives". For this reason, the evaluation of the model capabilities can be carried out only through the whole set of metrics.

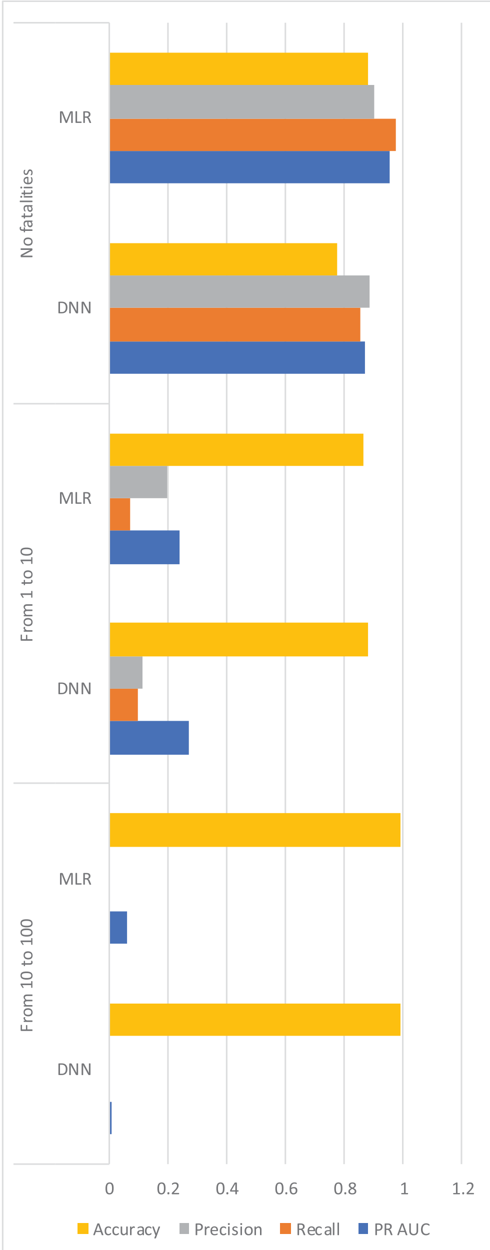


Fig. 2. Test results for the multiple linear regression (MLR) and deep neural network (DNN) models.

PR AUC gives a more complete overview of the model as it may indicate also the potential for improvement. In fact, it indicates the possibility to adjust the decision probability threshold to improve precision or recall based on the purpose of the analysis. For instance, tuning the threshold to optimize the model precision would be effective when predicting a rather frequent event with relatively low criticality, such as a hazardous event without fatalities. This would allow obtaining a model that tends to avoid false positives further increasing the number of alarms with some false ones. On the other hand, when predicting a rare event with relatively high criticality (such as a high-impact low-probability accidents (Pasman, 2015)), false alarms may be tolerated in exchange of a better prediction of the actual hazardous event, thus an improvement in true positive. For this reason, the threshold may be adjusted to optimize the recall.

A comparison between the two models shows a relatively better performance in the predictions obtained by MLR. However, if we compare the precision values for the severity category without fatalities, DNN turns to be more appropriate. Analogously, comparing the recall values for the severity category with fatalities between 10 and 100 shows a relatively better performance by DNN.

This demonstrates that there are some important differences among the specific techniques. Linear models such as MLR are widely used for prediction purposes. Interactions of the event features can be easily memorized through the provided datasets. However, a relatively simple model may not be able to capture the essential pattern in the data (Christian and Griffiths, 2016). Generalization of lessons learned for prediction under unknown circumstances requires a higher level of complexity, which linear functions may fail to provide (Goodfellow et al., 2016). Deep neural networks may be an option for such task (Christian and Griffiths, 2016).

Major accidents are (fortunately) rare events in industry, even considering evidence of fat-tailed distributions (Taleb, 2007). For this reason, appropriate models should be used to deal with such unexpected events. To this end, linear regression techniques are well-known for their limitation to handle rare events data (King and Zeng, 2001). Relatively simple models tend to forecast the basic trend and may potentially miss several exact points (Christian and Griffiths, 2016). Sophisticated models such as DNN are theoretically better suited to consider rare events, due to their sensitivity to input data and capability to generalize (Cheng et al., 2016). However, a limitation of DNN is that the model may have such a sensitivity to input data that the solutions it produces are highly variable (Christian and Griffiths, 2016). There can be errors in how the data were collected or reported on MHIDAS. For this reason, cross-validating with a test dataset is essential. Moreover, DNN results can be altered by its random initialization of parameters before every training session. This has the potential to affect the whole model development and, in turn, lead to slight alterations of prediction capabilities. Such differences may be amplified in case of relatively small datasets and few iterations to minimize the final loss function during training. Another limitation of the DNN model used in this case study may be related to its setting based on Cheng et al.'s (2016) work. In fact, the DNN model used may still need appropriate optimization for the case study.

An important aspect to consider is that the DNN model is not tied to a rigid structure to aggregate information from indicators (Landucci and Paltrinieri, 2016), but it has the potential to reshape its own structure based on new batches of data. Such an approach has some similarity with other methodologies in literature (Paltrinieri et al., 2016, 2015b), who developed a technique to update logic trees describing accident scenarios dynamically, in order to account for new evidence and prevent emergence of atypical events.

An option for improving the DNN model is the application of progressive learning techniques, which may be independent of the number of indicator categories and to learn new indicators once relevant information emerges, while retaining the knowledge of previous ones (Venkatesan and Er, 2016).

6. Conclusions

Through this work, we have suggested a data analytics approach for severity prediction of future hazardous events. The approach was based on the use of two well-known machine learning techniques: MLR and DNN. These models were developed and tested on the Major Hazardous Incident Data Service (MHIDAS) database. Part of the available data were used to build the actual models while the remaining data were used to test the models. This allowed also identifying and discussing the inherent limitations of the techniques.

For instance, DNN high model sensitivity does not tolerate inaccurate inputs. For this reason, selection and customization of a prediction model for an intended purpose should be carefully carried out using appropriate metrics, tolerance, and criteria. If these precautions are considered, the odds to deliver appropriate support for safety-related decision-making will be boosted. In this way, it will be possible to extract important information from heterogeneous data and effectively support safety-related decision making.

Acknowledgement

We thank Riccardo Solini for his work on this topic, which ultimately inspired this contribution.

References

AEA technology - Major hazards assessment unit, 2003. MHIDAS – Major Hazard Incident Data Service. UK.

Andrews, D.F., 1974. A Robust Method for Multiple Linear Regression. Technometrics 16, 523–531. https://doi.org/10.1080/00401706.1974.10489233

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., 2016. Wide & deep learning for recommender systems, in: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. ACM, pp. 7–10.

Christian, B., Griffiths, T., 2016. Algorithms to live by: the computer science of human decisions. Macmillan.

European Parliament and Council, 2012. Directive 2012/18/EU of 4 July 2012 on the control of major-accident hazards involving dangerous substances, amending and subsequently repealing Council Directive 96/82/EC - Seveso III. Off. J. Eur. Union 1–37.

Goodfellow, I.J., Bengio, Y., Courville, A., 2016. Deep learning, The MIT Press. Citeseer, Cambridge, Massachusetts, US.

Google LLC, 2018. Tensorflow [WWW Document]. Tensorflow TM. URL www.tensorflow.org

Hastie, T., Tibshirani, R., Friedman, J., 2009. Unsupervised learning, in: The Elements of Statistical Learning. Springer, pp. 485–585.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Polit. Anal. 9, 137–163.

Landucci, G., Paltrinieri, N., 2016. Dynamic evaluation of risk: From safety indicators to proactive techniques. Chem. Eng. Trans. 53. https://doi.org/10.3303/CET1653029

Øien, K., 2001. A framework for the establishment of organizational risk indicators. Reliab. Eng. Syst. Saf. 74, 147–167. https://doi.org/10.1016/S0951-8320(01)00068-0

Paltrinieri, N., Comfort, L., Reniers, G., 2019. Learning about risk: Machine learning for risk assessment. Saf. Sci. 118, 475–486. https://doi.org/https://doi.org/10.1016/j.ssci.2019.06.001

Paltrinieri, N., Cozzani, V., Wardman, M., Dechy, N., Salzano, E., 2010. Atypical major hazard scenarios and their inclusion in risk analysis and safety assessments, in: Reliability, Risk and Safety: Back to the Future. pp. 588–595.

Paltrinieri, N., Khan, F., 2016. New Definitions of Old Issues and Need for Continuous Improvement, in: Dynamic Risk Analysis in the Chemical and Petroleum Industry: Evolution and Interaction with Parallel Disciplines in the Perspective of Industrial Application. https://doi.org/10.1016/B978-0-12-803765-2.00002-0

Paltrinieri, N., Khan, F., Cozzani, V., 2015a. Coupling of advanced techniques for dynamic risk management. J. Risk Res. 18, 910–930. https://doi.org/10.1080/13669877.2014.919515

Paltrinieri, N., Øien, K., Tugnoli, A., Cozzani, V., 2013. Atypical accident scenarios: From identification to prevention of underlying causes, Chemical Engineering Transactions. https://doi.org/10.3303/CET1331091

Paltrinieri, N., Reniers, G., 2017. Dynamic risk analysis for Seveso sites. J. Loss Prev. Process Ind. 49. https://doi.org/10.1016/j.jlp.2017.03.023

Paltrinieri, N., Tugnoli, A., Bonvicini, S., Cozzani, V., 2011. Atypical scenarios identification by the DyPASI procedure: Application to LNG. Chem. Eng. Trans. 24, 1171–1176. https://doi.org/10.3303/CET1124196

Paltrinieri, N., Tugnoli, A., Cozzani, V., 2016. Dynamic Hazard Identification: Tutorial and Examples, in: Dynamic Risk Analysis in the Chemical and Petroleum Industry: Evolution and Interaction with Parallel Disciplines in the Perspective of Industrial Application. https://doi.org/10.1016/B978-0-12-803765-2.00004-4

Paltrinieri, N., Tugnoli, A., Cozzani, V., 2015b. Hazard identification for innovative LNG regasification technologies. Reliab. Eng. Syst. Saf. 137. https://doi.org/10.1016/j.ress.2014.12.006

Pasman, H.J., 2015. Risk Analysis and Control for Industrial Processes - Gas, Oil and Chemicals: A System Perspective for Assessing and Avoiding Low-Probability, High-Consequence Events. Elsevier Science.

Paté-Cornell, E., 2012. On "Black Swans" and "Perfect Storms": Risk Analysis and Management When Statistics Are Not Enough. Risk Anal. 32, 1823–1833. https://doi.org/10.1111/j.1539-6924.2011.01787.x

Samuel, A.L., 1959. Some Studies in Machine Learning Using the Game of Checkers. IBM J. Res. Dev. 3, 210–229. https://doi.org/10.1147/rd.33.0210

Solini, R., 2017. Data analytics for chemical process risk assessment: learning lessons from past events towards accident prediction. Bologna, Italy.

Sornette, D., 2009. Dragon-Kings, Black Swans and the Prediction of Crises. ETH Zurich, Chair of Systems Design.

Svozil, D., Kvasnicka, V., Pospichal, J., 1997. Introduction to multi-layer feed-forward neural networks. Chemom. Intell. Lab. Syst. 39, 43–62. https://doi.org/10.1016/S0169-7439(97)00061-0

Taleb, N., 2007. The black swan : the impact of the highly improbable. Random House, New York.

Venkatesan, R., Er, M.J., 2016. A novel progressive learning technique for multi-class classification. Neurocomputing 207, 310–321.

Article VI - Automated Power Lines Vegetation Monitoring using High-Resolution Satellite Imagery

Gazzea, M., Pacevicius, M., Dammann, D.O., Sapronova, A., Lunde, T.M., Arghandeh, R., 2021. **Automated Power Lines Vegetation Monitoring using High-Resolution Satellite Imagery**. Trans. Power Deliv. 1–9.

Article VII - Managing Heterogeneous Datasets for Dynamic Risk Analysis of Large-Scale Infrastructures

Pacevicius, M., Ramos, M., Roverso, D., Thun Eriksen, C., Paltrinieri, N., 2022a. **Managing Heterogeneous Datasets for Dynamic Risk Analysis of Large-Scale Infrastructures**. Energies. p. 40.

energies

MDPI

Article

Managing Heterogeneous Datasets for Dynamic Risk Analysis of Large-Scale Infrastructures

Michael Felix Pacevicius [1,2], Marilia Ramos [3], Davide Roverso [2], Christian Thun Eriksen [4] and Nicola Paltrinieri [1,*]

1 Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, Richard Birkelands vei 2B, 7034 Trondheim, Norway; michael.pacevicius@esmartsystems.com
2 Analytics Department, eSmart Systems, Håkon Melbergs vei 16, 1783 Halden, Norway; davide.roverso@esmartsystems.com
3 The B. John Garrick Institute for the Risk Sciences, University of California, Los Angeles (UCLA), Los Angeles, CA 90095, USA; marilia.ramos@ucla.edu
4 Architecture Development Department, eSmart Systems, Håkon Melbergs vei 16, 1783 Halden, Norway; christian.t.eriksen@esmartsystems.com
* Correspondence: nicola.paltrinieri@ntnu.no

Abstract: Risk assessment and management are some of the major tasks of urban power-grid management. The growing amount of data from, e.g., prediction systems, sensors, and satellites has enabled access to numerous datasets originating from a diversity of heterogeneous data sources. While these advancements are of great importance for more accurate and trustable risk analyses, there is no guidance on selecting the best information available for power-grid risk analysis. This paper addresses this gap on the basis of existing standards in risk assessment. The key contributions of this research are twofold. First, it proposes a method for reinforcing data-related risk analysis steps. The use of this method ensures that risk analysts will methodically identify and assess the available data for informing the risk analysis key parameters. Second, it develops a method (named the *three-phases method*) based on metrology for selecting the best datasets according to their informative potential. The method, thus, formalizes, in a traceable and reproducible manner, the process for choosing one dataset to inform a parameter in detriment of another, which can lead to more accurate risk analyses. The method is applied to a case study of vegetation-related risk analysis in power grids, a common challenge faced by power-grid operators. The application demonstrates that a dataset originating from an initially less valued data source may be preferred to a dataset originating from a higher-ranked data source, the content of which is outdated or of too low quality. The results confirm that the method enables a dynamic optimization of dataset selection upfront of any risk analysis, supporting the application of dynamic risk analyses in real-case scenarios.

Keywords: heterogeneous datasets; metadata; dynamic risk analysis; potential of knowledge; power grids

Citation: Pacevicius, M.F.; Ramos, M.; Roverso, D.; Eriksen, C.T.; Paltrinieri, N. Managing Heterogeneous Datasets for Dynamic Risk Analysis of Large-Scale Infrastructures. *Energies* **2022**, *15*, 3161. https://doi.org/10.3390/en15093161

Academic Editor: Petar Sarajcev

Received: 22 February 2022
Accepted: 13 April 2022
Published: 26 April 2022

1. Introduction

Electric energy plays a crucial role in today's society, and it is involved in almost all aspects of society's daily routine [1]. The continuous development of the economy increases the need for energy, leading to larger-scale power systems and increasingly complex structures [2]. Furthermore, the scale and complexity of power grids are expected to increase with the growing use of renewable energy sources [3], as well as the development and implementation of smart grids [4]. As numerous businesses, public infrastructures, and private households rely on the provision of power for their daily tasks, companies in charge of the power supply need to provide energy management in a more reliable, effective, and secure way [1,5].

Power grids are exposed to a plurality of hazards such as hurricanes, earthquakes, ice storms, and floods, which can have severe consequences. The increasing frequency of extreme weather events increases the damage potential of these hazards, further highlighting the vulnerability of power systems [6]. Indeed, large-scale power outages have frequently occurred in recent years and have caused substantial economic losses [2].

Risk assessment and management have received significant attention as a tool to ensure the operational safety and reliability of power systems, becoming one of the major tasks for urban power-grid management [2]. Risk assessment of power grids generally makes use of traditional risk approaches such as reliability block diagram (RBD), fault tree analysis FTA [7], or failure mode and effect analysis [1]. Yet, the complexity of the power grids and the growing amount of data coming, e.g., from prediction systems favor the development and use of more advanced probabilistic risk-based approaches [6]. Applications of data-based approaches to power grids and energy systems range from energy management for smart buildings [8] to online fault diagnosis [9], among others (e.g., [10,11]). Risk analysis of power grids susceptible to vegetation-related hazards can adopt diverse data sources, ranging from satellite-based orthophotos to drone-based aerial images, including plane-based orthophotos or LiDAR 3D point clouds. Connected devices and access to more computing power provide additional opportunities for data-based, dynamically updated risk analyses. However, an updated and accurate risk analysis is highly dependent on the data used to inform the different parameters for calculating risk, e.g., the frequency of an event, the probability of failure, and the potential consequences of this event. Indeed, the use of different datasets for failure frequencies is an important source of uncertainty in risk analysis results [12].

In addition to better informing conventional risk analysis, multiple data sources present an opportunity for dynamic risk analysis (DRA). DRA is a concept that has mostly emerged over the last decade [13,14]. It aims to build on data availability to provide more frequent and performant risk pictures of infrastructures [15]. While DRA can benefit from the growing data source variability to diversify the possibilities of information acquisition relative to a particular parameter [4,16,17], numerous data sources may also increase data collection and processing complexity [4,18]. First, the data to process are intrinsically heterogeneous, requiring a large panel of competencies to manipulate and extract relevant information from the datasets. Second, a larger number of data sources requires selection rules for decision-making optimization, given the potential variability in the data quality. This variability can be due to, for instance, the type of considered datasets, the spatiotemporal resolution of the data, or the acquisition conditions of the datasets.

The International Standard Organization states that risk assessment should use the "best information available" and the implementation of "dynamic" approaches [19]. However, there is no guidance for applying those principles when multiple data sources are available. The present paper is a step toward closing this gap. We propose an approach for the dynamic optimization of dataset management to reduce uncertainties relative to data selection upfront of any risk calculation. The proposed method (called the three-phases method) is based on metrology concepts and metadata for characterizing the parameter-related information needed for a quantitative risk analysis (QRA).

The method focuses on three main features of the datasets impacting the quality and usability of the data for a QRA: the nature of a dataset, the discrepancies observed between the spatiotemporal attributes of the dataset and the spatiotemporal requirements for the risk analysis, and the agents and factors involved in the data management. The method integrates these three factors in a scoring system using meta-features, relying solely on metadata. The result is a ranking of the datasets, based on their informative potential relative to a baseline of "perfect information". The method also predicts the informative potential of any new dataset originating from a list of preselected data sources using only the information available in the metadata, thus without factually analyzing the content of the datasets. Hence, the method's application allows a continuous selection of the best candidate across all available datasets. While the implementation of the method is labor-intensive in the first iterations of the process, it can significantly increase data management efficiency in future assessments in the long term, avoiding suboptimal repetition of tasks.

The application of the method is demonstrated through a case study focusing on risk management in power grids. We focus on the role of vegetation along power lines, which represents a common source of outages in power grids, either via trees falling on the power lines or by growing under the infrastructure until grounding one phase [4,5,20–22].

The remainder of this paper is organized as follows: Section 2 describes the methods and concepts on which the three-phases method is founded: metadata and risk analysis/dynamic risk analysis. Section 3 presents the result, i.e., the method developed for dataset management on risk analysis. It describes the preliminary actions required for the application of the three-phases method, which is then fully detailed in the rest of the section as the main contribution of this work. Section 4 presents an application of the method to the case study. Section 5 discusses the case study results, as well as the benefits and limitations of the method, followed by conclusions in Section 6.

2. Materials and Methods—Metadata and Risk Analysis

Metadata can be defined as "data that provide information about other data" [23]. Metadata can provide structured information about a dataset without analyzing the dataset content. As highlighted by Wierling et al. [24], credible and traceable documentation of knowledge about the energy system is not possible without metadata. Despite its potential benefits for energy systems and data management optimization, the assessment of datasets through their metadata is not extensively explored in risk analysis. Indeed, there is no uniform definition of metadata to standardize the entire process of data production, processing, analysis, and use for prediction in the field of safety [25].

Data source management using its metadata in the context of risk analysis requires, at first, a clear definition of the level of analysis. In addition, it requires having a complete picture of all the datasets usable to inform the risk analysis parameters (i.e., an exhaustive description of the risk analysis parameters and a list of all the data sources usable to inform those parameters). While these two actions are common steps in risk analysis, they generally lack details that would enable an optimal dataset management. A reinforcement of those steps ("reinforcement actions") is, thus, needed, as introduced in Section 2.2 and further detailed in Section 3.

This section presents an overview of concepts related to metadata and risk analysis. These do not constitute an exhaustive review and are limited to the description of the concepts applied in this paper.

2.1. Metadata Concepts

Metadata (i.e., "data that define and describe other data" [26]) report information concerning the structure and the content of a dataset or a service [27,28]. Metadata can be used for three main purposes: (1) content description (author, subject, etc.), (2) structural characterization (e.g., link between various parts of a resource), and (3) administrative management (access rights, file version, etc.) [29]. In addition to these features, metadata can be classified on the basis of a piece of information's intrinsic vs. extrinsic property [30,31]. Although intrinsic properties may be assimilated to (1) content description and extrinsic properties cover, (2) structural characterization, and (3) administrative management, there is no broad consensus on the topic [32,33]. The classification and the metadata quality assessment depend, thus, on the task at hand [34], leading to new classifications if required.

Different metadata standards have been developed over the years, depending on the fields of application and the metadata's purposes. The Metadata Standards Directory Working Group [35], a working group from the Research Data Alliance [36], has reported a community-maintained "open directory of metadata standards applicable to scientific data" [37]. An extract of this work is presented in Appendix A. This directory also reports the Dublin Core (DC), which is a generic standard developed on Semantic Web principles (or a "web of linked data") [38,39] and managed by the Dublin Core™ Metadata Initiative, or DCMI. DCMI aims at developing and sharing best practices in the design and management of metadata. It is an open, collaborative, international, cross-disciplinary, technology-neutral, and business model-/purpose-neutral organization dedicated to supporting metadata management since the 1990s [40,41]. Dublin Core is a widely used standard, also published as an ISO standard and NISO standard [42–44]. It contains 15 core terms and several properties, classes, datatypes, and vocabulary encoding schemes maintained by DCMI (DC terms) [45].

The adoption of the DC standard data management for risk analysis presents several advantages, such as the following:

- many of the data sources not conventionally considered may be made available online,
- cross-disciplinary standards are critical to the comparison of heterogeneous data sources,
- the importance taken over the years and continuous increase in cloud-based technologies and web-based applications,
- the importance of facilitating the sharing of data and knowledge, the collaboration, the research and development, and the innovation adoption to third parties both in the risk community and across industries.

Furthermore, using the DC standard allows using DC-related crosswalks, facilitated by the international long-term recognition of the DC metadata standard. Crosswalks enable highlighting the nature of the overlap and gaps between different metadata standards through a table or a figure. In addition, they allow pinpointing the existence or the absence of relationships between terms existing in the respective standards [46]. Multiple examples of crosswalks linking recognized schemata can be found online, such as the one provided by the Getty Research Institute [47], the one provided by the Metadata Working Group of the Emory University [48], or the one provided by the UBC Faculty Research and Publications [49]. Non-standardized crosswalks (e.g., internal) may also be considered when discrepancies are observed between the format followed for metadata reporting in a selected file and the existing standards (e.g., due to explicit choices related to specific metadata needs, or due to a simple lack of competencies). Hence, the content from other schemas can always be linked to the Dublin Core schema.

2.2. Conventional Risk Analysis and Dynamic Risk Analysis

The concept of risk is generally related to three principal elements, as displayed in Equation (1) [50].

$$\text{Risk} = \text{f}(s, p, c), \tag{1}$$

where s corresponds to a specific scenario, p corresponds to the probability of occurrence of this specific scenario, and c corresponds to the resulting consequences.

Various standards adopt this definition for defining the steps of risk assessment [19,51,52]. Figure 1 presents the different steps to be followed within a risk assessment [51]. The figure also presents the placement of the proposed reinforcement actions to be described in Section 3.1, in blue and in red. Note that additional steps are identified in green in the figure: the management of datasets for informing the risk assessment. These steps consist of the method proposed in this paper, as described in Sections 3.2 and 3.3.

Different sources of uncertainties may arise during a risk assessment, one of them being data processing [53]. The "level of knowledge" to represent some of these sources was added to Kaplan and Garrick's definition (Equation (2)) by Aven and Krohn [54].

$$\text{Risk} = \text{f}(s, p, c, k), \tag{2}$$

where the variable k corresponds to the level of knowledge and is added to the variables s, p, and c corresponding to scenario, probability of occurrence, and consequence, respectively.

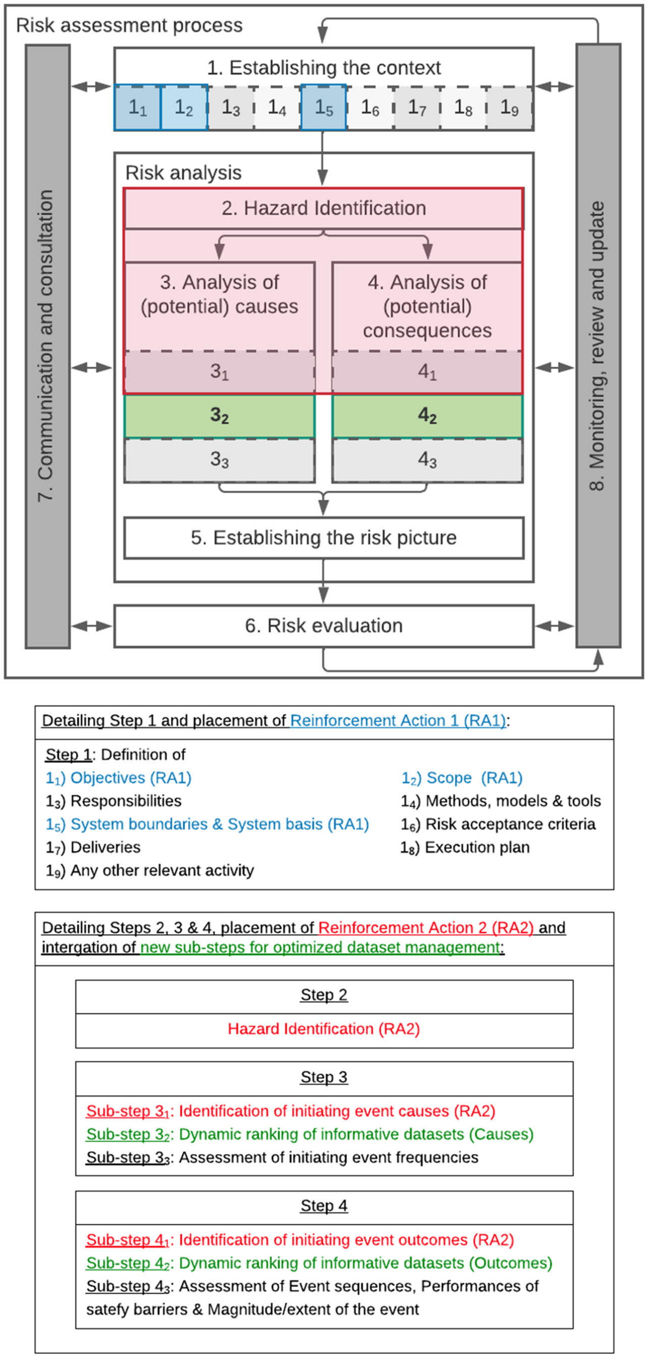


Figure 1. Augmented risk assessment framework Z-013 [51]. The steps highlighted in blue and red are the subject of augmentation (reinforcement actions 1 and 2), and the steps 3_2 and 4_2 in green are additional steps related to the optimization of data source/dataset management.

The assessment of the level of knowledge requires a proper characterization of the information pipeline, starting with the data acquisition [16,19]. The concept of "best level of information" selection, associated with the concept of "dynamicity", can help in ensuring more efficient risk assessment and having a clear picture of the related uncertainties.

The notion of dynamicity was recently added to the principles of risk management presented within ISO 31000. Dynamic risk management approaches aim not only to update the data to consider, but also to adapt and reconsider, if necessary and on the basis of new risk evidence [55], the assumptions and models retained in previous cycles of the assessment [15,56–58]. As such, those techniques avoid lock-ins from initially considered conditions and process inertia by integrating, by design, the possibility to appropriately reshape the risk assessment process while minimizing the required efforts [59].

Despite the increasing number of publications and recognition of its relevance in ISO 31000:2018, DRA remains in an embryonic phase [15,60,61]. Limited research in the field hinders its implementation and the possibilities of improvements of DRA techniques. The lack of a systematic approach for identifying available data, as well as characterizing and managing data sources, also poses a challenge for the adoption of DRA, as it is a data-driven method. The method proposed in this paper intends to address this gap through the reinforcement actions detailed in Section 3.1 and the addition of two steps, presented in Sections 3.2 and 3.3.

3. Results—Dataset Management Method for Dynamic Risk Analysis of Large-Scale Infrastructures

This section presents the resulting method developed for dataset management. It starts by describing the reinforcement steps required to apply the three-phases method. In Section 3.1, the main building blocks of the method are presented in Section 3.2, followed by the detailed description of the method elements in Section 3.3.

3.1. Risk Analysis Framework Reinforcement: Level of Analysis and Dataset Characterization

This subsection first details to which extent information should be characterized to enable a standard risk assessment. It then presents two reinforcement actions (RA1 and RA2) applied to existing steps of a standard risk assessment (Figure 1), namely, establishing the context (sub-steps 1_1, 1_2, 1_5) (RA1) and hazard identification (step 2), analysis of potential initiating event (sub-step 3_1), and analysis of potential consequences (sub-step 4_1) (RA2). The reinforcement of these steps is necessary for applying the proposed method for dataset selection (Sections 3.2 and 3.3).

3.1.1. Information Characterization Requirements

Considering that the numerical values used within a QRA are all directly or indirectly based on measurements, best practices applied in metrology (i.e., the "science of measurement and its application" [62]) can be adopted as a reference. The measurement process in metrology is defined as "a set of operations to determine the value of a quantity" [63]. Its design represents a critical phase and consists, from a high-level point of view, in answering the following questions to execute a measurement adequately:

- Which quantity shall be measured?
- What are the required quality indicators (e.g., accuracy, precision, (see Figure 2))?
- Which measurement methods shall be used?
- Which equipment shall be used?
- Which software shall be used?
- Who is going to execute the measurement?
- What are the ambient conditions and influencing quantities affecting the measurement process?

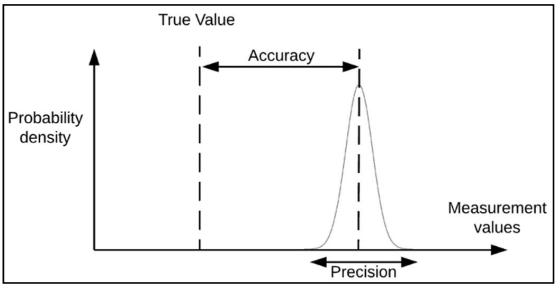


Figure 2. Accuracy–precision distinction. Illustration of the distinction between the concepts of accuracy (closeness of agreement between a measured value and a sought out true value) and precision (closeness of agreement between measured values obtained by replicate measurements on the same or similar objects under specified conditions) [62].

Providing the described level of detail is critical for the validity of a measurement result, and to improve the traceability of a measurement. This is particularly relevant for risk assessment and recalls the paramount importance of a proper context characterization. Indeed, answering the question "Which quantity shall be measured?" requires first an adequate identification of the information that is sought out. This action should be executed within step (1) of the risk assessment (Figure 1) ("establishing the context"), as part of the global definition of the problem to address.

Three main points among those reported in the context establishment of the NORSOK Z-013 standard [51] need to be defined to adequately characterize the information one should look for:

- The objectives (defining the objective functions and indicating which type of information should be chosen),
- The scope (characterizing to which extent this information needs to be researched),
- The system boundaries (characterizing under which considerations and within which system delimitations the data need to be sought out).

3.1.2. Reinforcement Actions: Level of Analysis and Available Data Sources

The implementation of risk analyses is, in practice, strongly constrained by the availability of needed resources [64,65]. Hence, the adequate level of analysis is a tradeoff decision between stakeholder expectations and analytical possibilities [66,67]. Figure 3 illustrates the nature of the tradeoff to be found when defining the optimal level of analysis. The optimal analysis level can be considered as the level of convergence between a top-down and a bottom-up process. The top-down process consists of the progressive detailing of a global ambition associated with a resource budget allocation. The bottom-up process consists of progressively aggregating and restoring required information most efficiently while reducing information loss [68]. The dotted line in Figure 3 can be read as the level of convergence; it can be scrolled up or down depending on objectives and conditions. Note that no budget would enable a microscopic analysis of a large and complex system, and some level of abstraction is inevitable. On the other hand, no analysis can be limited to a high-level identification of risk-related objectives, and some level of details will always be required for meaningful decision making.

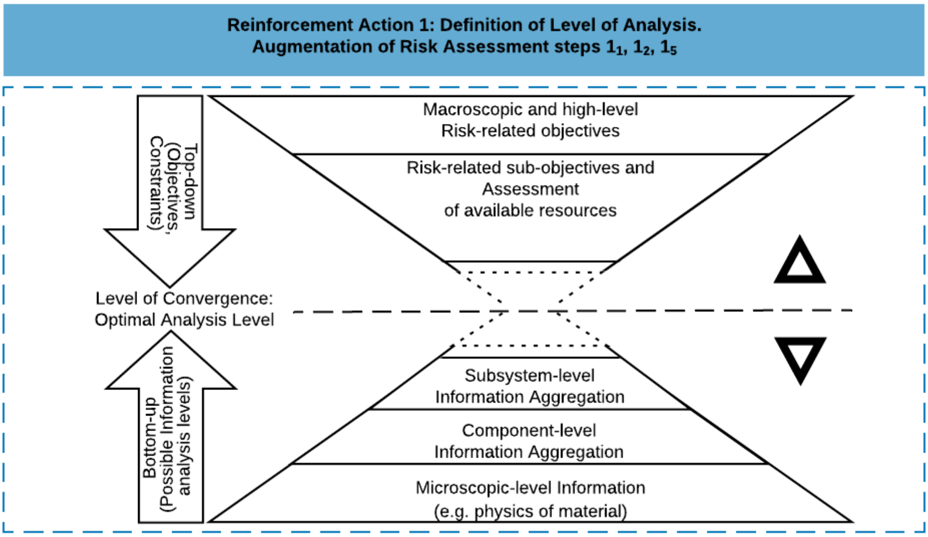


Figure 3. Level of analysis of a risk assessment defined as tradeoff decision between stakeholder expectations and analytical possibilities.

The definition of the optimal level of analysis is related to steps 1_1, 1_2, and 1_5 of the risk assessment method, namely, definition of objectives, definition of the scope, and definition of system boundaries and systems basis (Figure 1). Defining the level of convergence (reinforcement action 1) supports identifying the most relevant system, subsystem, assembly, subassembly, or component on which a risk assessment shall be executed. Following this identification, the next phase consists of building on the following steps commonly applied in risk assessment [69], augmented with reinforcement action 2, as presented in Figure 4:

- Task (I): applying a hazard identification (HAZID), i.e., identifying all relevant hazards and hazardous events,
- Task (II): describing the relevant accident scenarios,
- Task (III): reporting all dimensions to be considered for the hazardous events addressed in each scenario, from both a probability and a consequence perspective,
- Task (IV): identifying and characterizing all relevant parameters per reported dimension,
- Task (V): identifying all the data sources providing, to any extent, information to those parameters on the basis of experience, expertise, and further benchmarks.

Reinforcement action 2 consists, thus, of preselecting a list of data sources to inform different parameters which, in turn, inform different dimensions needed for quantifying the probability of occurrence and the consequences of a specific scenario. The list of preselected data sources should be completed by looking at all the accessible data sources and determining if those can provide (to any extent possible) knowledge about the needed parameters. For traceability, the preprocessing tasks enabling one to link a data source to a parameter shall also be reported.

The characterization of the parameters (Task IV) is a crucial step. It starts by reporting attributes relevant in any measurement process, i.e., the unit, the optimal resolution, and the range.

At this stage, considering that suboptimal resolution may often be faced, it is also strategical to define acceptable subcategories of information as second-best options to enable a semiquantitative evaluation when no other possibilities exist.

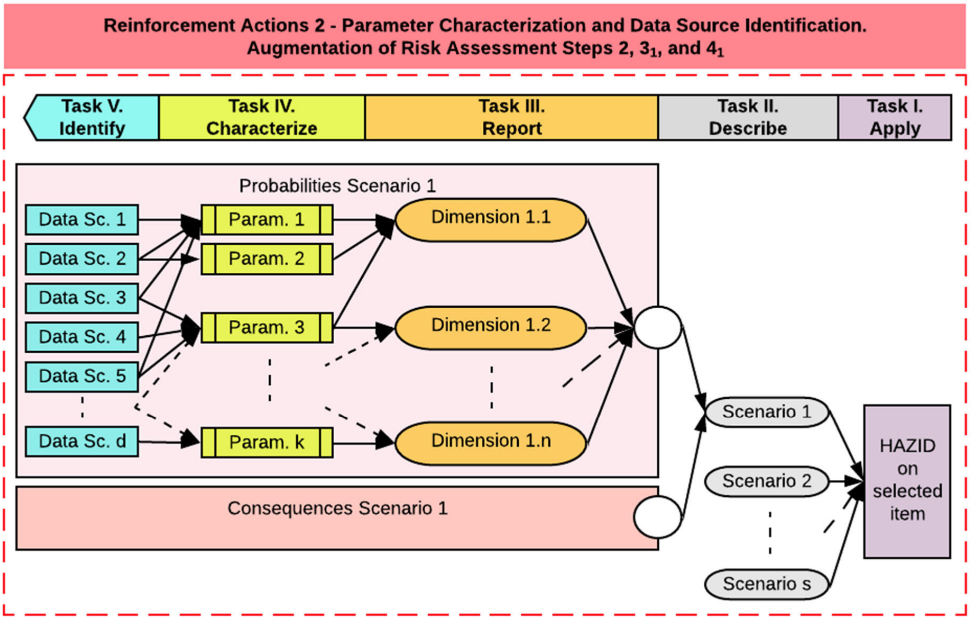


Figure 4. Identification of parameter characterization requirements and data sources potentially exploitable for the risk analysis.

Risk analysis should ideally be site-specific [13,58] and performed in real time to avoid the possibility of building on outdated data and outdated considerations [70]. Therefore, risk analyses are time- and location-sensitive, and any spatiotemporal divergence between the site to be analyzed and the data that are considered will impact the results. Thus, the following questions should also be answered for characterizing the risk parameters:

- How location-sensitive is the parameter under review?
- What is the spatial extrapolation potential, i.e., the capacity, given data provided for a particular parameter in a delimited geographical area, to estimate values for that specific parameter in the surrounding of the initially considered area?
- How quickly does the parameter under review usually change over time?
- What is the relevant time changing rate?
- How long would it take before the dataset considered for the parameter under review to be outdated?

Depending on the scope of the risk assessment being performed, an applicable spatial scale may be the following (in square meters): "not applicable (NA), individual or $<10^0$", "$<10^1$", "$<10^2$", "$<10^3$", "$\geq 10^3$". Similarly, a timescale could be reported as "hours", "days", "weeks", "months", "years", "decades", or "constant" (i.e., no change over the lifetime of the site).

In summary, a parameter *pa* can be characterized through the vector

$$pa = \begin{bmatrix} Rs_{pa} \\ SLI_{pa} \\ Ra_{pa} \\ SEP_{pa} \\ TS_{pa} \end{bmatrix},$$

where, within a pre-defined scope, Rs_{pa} corresponds to the optimal resolution of the parameter pa based on the chosen unit, SLI_{pa} corresponds to the sublevel of information of the parameter pa acceptable for semiquantitative evaluations, Ra_{pa} corresponds to the range of values taken by the parameter pa, SEP_{pa} corresponds to the spatial extrapolation potential of the parameter pa, and TS_{pa} corresponds to the temporal sensitivity of the parameter pa.

Thus, the implementation of the actions reported up to step 3_1 in Figure 1, reinforced with the reinforcement actions 1 and 2, allows obtaining a preselection of all potentially relevant data sources. Additionally, it enables one to precisely list the attributes usable for a quality assessment of the information provided by a dataset in terms of risk quantification.

3.2. Dataset Management: Three-Phases Method Overview

Data quality assessment has a long research history [71] and is usually executed by comparing the value of specified data quality indicators to preliminary defined reference values. The quality of the information can be assessed using various dimensions, such as accuracy, precision, coverage, completeness, timeliness, reliability, trustworthiness, traceability, comparability, costs, and metadata [72–79]. Section 3.2.1 discusses the most relevant dimensions for risk analysis and shows how those can be characterized using the terms defined in the Dublin Core standard. This is then used as the foundation for the proposed data management method, described in Section 3.2.2.

3.2.1. Dataset Characterization for Risk Analysis

Efficient dataset management for risk analysis relies on the characterization of three main features, as described below: nature of the dataset, site/time specifications of the dataset, and agents and factors influencing data management.

(*i*) *Nature of the dataset*
The technologies used to capture data determine which type of file will be generated. This directly impacts the obtainable performance in terms of resolution, range coverage percentage (how much of the predefined range can be covered), precision, and accuracy. For instance, the best spatial resolution available via commercial satellite images is much lower than that provided by LiDAR point clouds (30 cm vs. a few millimeters) [80–82]. Furthermore, satellite images are mainly used to provide 2D information, while LiDAR point clouds are usually used to obtain 3D insights.

(*ii*) *Spatiotemporal characterization of data*
Figure 5 illustrates information provided for a unique and generic parameter, at three different resolutions, at three points in time (t − 2, t − 1, and t), for a specific area of interest (*AoI*). While the most recent dataset with the highest resolution would be ideal, datasets are most often incomplete. Therefore, one may face situations where the highest spatial resolution is only available within an older dataset (e.g., t − 2 here), making datasets with coarser spatial resolution the only up-to-date option [83]. Additionally, one may also face a total absence of information in some regions (represented by the black region).
The management of incomplete datasets is an important task to be performed for most of the parameters involved in a risk analysis. This highlights the importance of adequately addressing the spatiotemporal characterization of the information provided by a dataset, and including it as a comparison and evaluation criterion.

(*iii*) *Agents and factors influencing data management*
The value of information available in a dataset strongly depends on the competencies of the actors involved in the various steps of the data management (i.e., data capture, data transmission, data storage, data pre-processing, information processing, results transmission) [16,71]. The trust to be given to the information provided by a dataset is, thus, strongly influenced by, e.g., the standards and protocols followed when managing the data, the authority, and legitimacy of the actors involved [39,84].
Identifying the "trust" level, the spatiotemporal features and the nature of the dataset are, thus, essential for the characterization of the datasets to be used for risk as-

sessment. These three features are the foundation for the data management in the three-phases method. Note that the implementation of reinforcement actions 1 and 2 as previously described is required to apply the method (Figure 6).

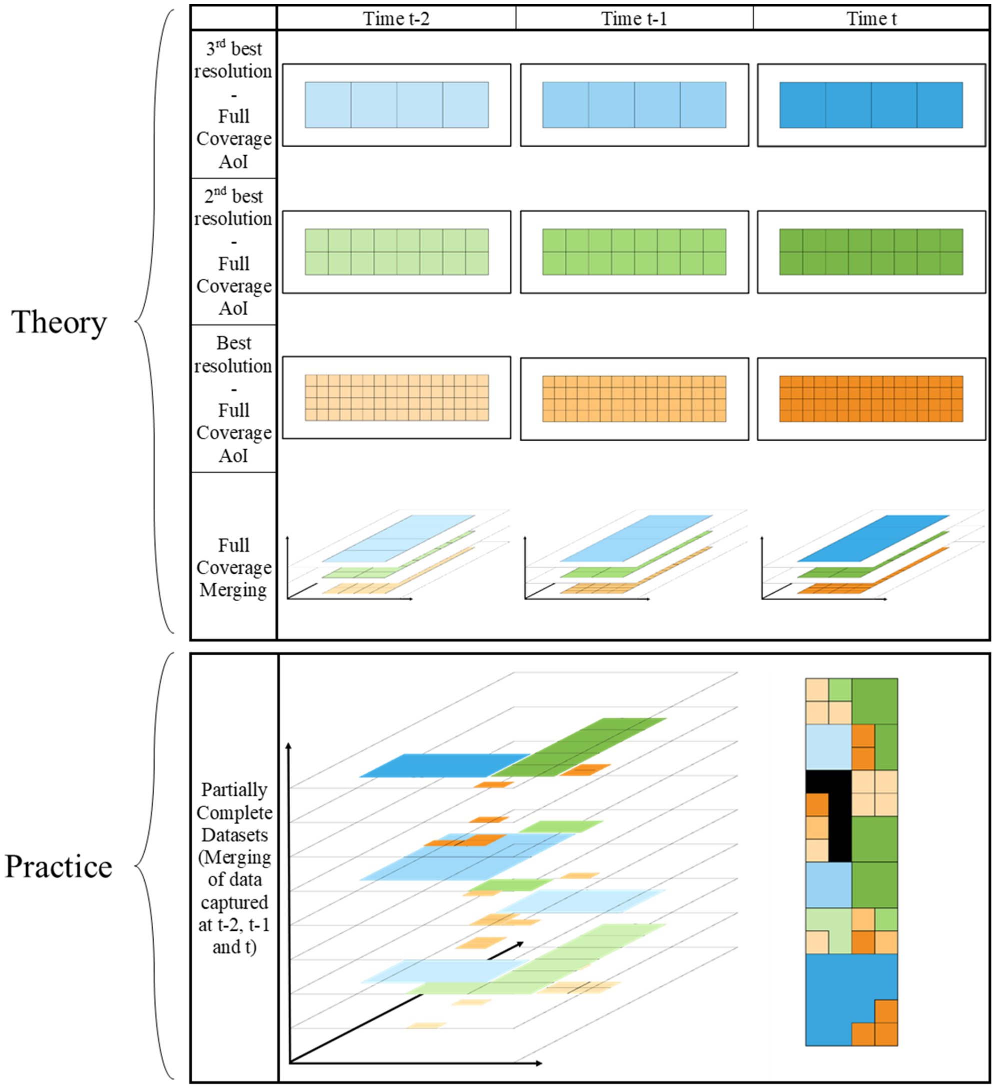


Figure 5. Differences between theory and practice in the spatiotemporal characterization of a generic parameter in a predefined area of interest (*AoI*). Incomplete datasets encountered in practice lead to the dilemma of sometimes having to choose between resolution and timeliness to inform the chosen parameter. Additionally, some regions may show total absence of information (black region in merged 2D view of the *AoI* at the bottom right of the image), which is particularly problematic for a risk assessment to be executed in that area.

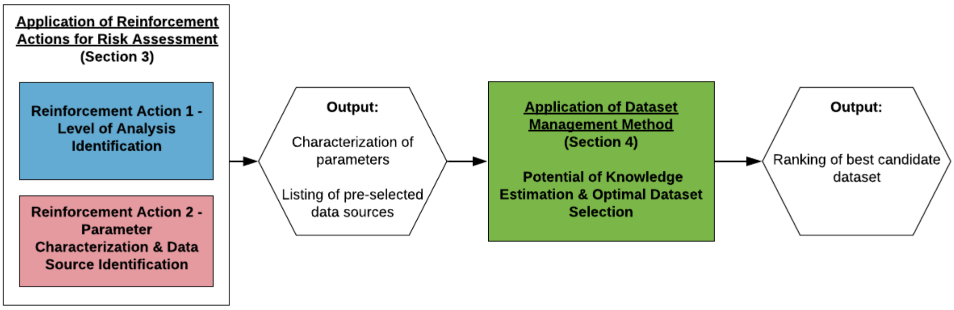


Figure 6. Logic of steps for applying the data management method. The reinforcement actions applied on common risk assessment frameworks provide the parameter characteristics required for a QRA, as well as a list of data sources that can be used to inform those.

3.2.2. Three-Phases Method—Logic Description

The Dublin Core standard presented in Section 2.1 is used as a foundation to exploit the metadata in the three-phases method. We start by only selecting the terms that are relevant for risk assessment purposes, i.e., those related to the three features defined in Section 3.2.1. We then regroup the terms into three classes by following a similar logic: (1) *file* (nature of the dataset), (2) *scene* (site-/time-specifications of the dataset), and (3) *objectives/author/circumstances* (agents and factors influencing data management). Tables A2–A4 in Appendix B detail this recategorization, together with the respective definition of each of the selected terms [45].

The terms categorized in the first class ((1) *file*) report the nature of the file. They are used to characterize the default maximum potential of knowledge (*DMPK*) that a specific data source can provide, based on the technological possibilities of the technique used to generate the dataset (e.g., satellite-based orthophoto, LiDAR-based point cloud).

The terms categorized in the second class ((2) *scene*) report the spatiotemporal properties of the file. This class can be divided into two subclasses: (2a) *spatial* and (2b) *temporal*. The use of information provided in class (2) *scene* enables one to calculate a first degradation factor (DF_1, composed of DF_{1a} and DF_{1b}, relative to spatial and temporal information, respectively) on the basis of the difference in nature between the spatiotemporal requirements of the site to be analyzed and the spatiotemporal properties of the considered dataset.

The terms categorized in the third class ((3) *objectives/author/circumstances*) report contextual information. They enable calculating a second degradation factor (DF_2), characterizing the level of trust one assigns to the analyzed dataset. In opposition to the first degradation factor, the second degradation factor calculation can be considered as a more dynamic and subjective task, as the trust level is strongly influenced by the stakeholders supervising the risk analysis [85]. For instance, understanding a problem and the knowledge of the mentioned actors/standards could be very different between two distinct teams [86], a standard may become outdated and withdrawn after some time, etc.

Figure 7 illustrates the sequencing of the phases required to calculate an assessed dataset's informative potential (i.e., potential of knowledge).

The notions of "degradation factors" have been chosen because divergences observable via the analysis of properties relative to terms in class (2) *scene* and (3) *objectives/author/circumstances* can only neutrally or negatively impact the maximum performances of the knowledge acquired via the analysis of properties relative to terms in class (1) *file*.

The analytical order of the phases aims to optimize future data processing: the spatial overlap is assessed before analyzing the temporal properties to automatically discard non-overlapping datasets. Furthermore, one may still decide to valorize the analysis of properties relative to terms present in classes (1) *file* and (2) *scene*, despite a lack of qualifications leading to an absence of trust-related quality check. Including trust-related quality checks in the final calculation may, thus, have to be appreciated on a case-by-case study of the problems, justifying a final position for this task in the method.

Table 1 presents the assumptions considered during the development of the method, followed by a detailed description of each of the phases in the next subsection.

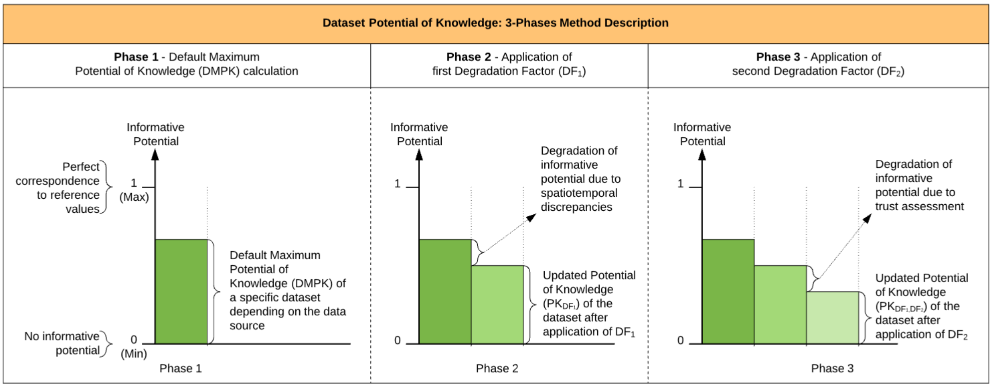


Figure 7. Three-phases method description. The figure shows the progressive degradation in the assessment of a dataset's informative potential when compared to the originally required level of information.

Table 1. List of assumptions made for the development of the presented method.

N°	Assumptions
1	We place ourselves in a situation where we can apply all steps previously discussed (i.e., optimization of analysis level, HAZID on selected item, scenario identification, probability, and consequence characterization (i.e., identification of all involved dimensions, parameters and usable data sources), characterization of the required information for each parameter, and ability to report the metadata of the selected datasets following the DC standard).
2	A consensus is assumed among all the stakeholders involved in each method development step.
3	All datasets are initially considered external to the stakeholders involved in the risk analysis, thus needing to go through the method similarly.
4	All analyzed datasets are considered independent.
5	All datasets are considered to be analyzed independently and not leveraging on one another.
6	The quality of the datasets analyzed in earlier risk analyses is assumed to be optimized regarding acquisition conditions and state-of-the-art possibilities in the field (resolution, scale calibration, etc.), and the data are considered to be acquired by an expert.
7	The metadata of all datasets are convertible in DC terms.
8	No advanced natural language processing is used to extract information from text in this first version of the method.
9	A data source can uniquely be identified on the basis of the format and the type of a resource.
10	The number of most obvious invalid records can be indicated using dataset quality indicators. Although not originally reported in the DC standard, such information can easily be added to existing metadata.
11	The number of missing values can be indicated using dataset quality indicators. Although not originally reported in the DC standard, such information can easily be added to existing metadata.
12	Missing values can be characterized in time and/or space (when relevant).
13	Trust-related properties are dataset-specific and generalizable to all parameters informed.
14	All datapoints of the same dataset are acquired using a unique acquisition process.
15	The reported spatiotemporal information of datasets is assumed to be accurate (no mismatch).

3.3. Quantitative Elements of the Three-Phases Method

The three-phases method identifies and characterizes multiple data source/dataset properties through a number of classes and respective boundaries. The characterization of these classes is based on the authors' experience with data management for risk analysis. Those choices are valid from a generic perspective to the best of our knowledge. However, the method offers the flexibility for this information to be adapted to the context in which the method will be applied. The values shall, thus, be seen as an indication instead of a static and rigid formalization. The implications of those choices are further discussed in Section 5.

3.3.1. Phase 1: Default Maximum Potential of Knowledge (*DMPK*)—Calculation

The evaluation approach of class (1) *file* consists of the calculation of the *DMPK*, which is assessed per data source *sr* and per parameter *pa* (i.e., $DMPK_{sr,pa}$). The $DMPK_{sr,pa}$ is a function of four properties identified on the basis of common data quality assessment criteria [71] to estimate how well preselected data sources can inform a parameter. The *DMPK* can be calculated through a normalized weighted sum as in Equation (3).

$$DMPK_{sr,pa} = \frac{x_{LoI}\,\frac{LoI_{sr,pa}}{3} + x_{RaC}\,\frac{RaC_{sr,pa}}{3} + x_{Pr}\,\frac{Pr_{sr,pa}}{2} + x_{Ac}\,\frac{Ac_{sr,pa}}{2}}{x_{LoI} + x_{RaC} + x_{Pr} + x_{Ac}}, \quad (3)$$

where the variables are defined as follows:

- $DMPK_{sr,pa}$: default maximum potential of knowledge per data source *sr* and per parameter *pa*,
- $LoI_{sr,pa}$: the level of information for source *sr* and parameter *pa*,
- $RaC_{sr,pa}$: the range coverage for source *sr* and parameter *pa*,
- $Pr_{sr,pa}$: the precision for source *sr* and parameter *pa*,
- $Ac_{sr,pa}$: the accuracy for source *sr* and parameter *pa*,
- x_{LoI}, x_{RaC}, x_{Pr}, x_{Ac}: weights given by stakeholders to the level of information, the range coverage, the precision, and the accuracy of the data, respectively.

The weights give stakeholders the possibility to manage the importance given to meta-parameters as wished. For simplicity, a naïve approach setting those weights to 1 is applied for the rest of the present paper [87].

The use of the *DMPK* enables a first ranking of data sources based on their capacity to inform a specific parameter. Thereby, any new dataset *ds* originating from one of the reported data sources will automatically be given a *DMPK* score enabling an estimation of its a priori value for risk analysis.

Calculating the *DMPK* allows the stakeholders to identify the parameter characterization benefiting the most from data coming from a specific data source by assessing the *DMPK* scores for a unique source, and identify which dataset shall be used to inform a particular parameter depending on the origin of the different sets.

The four properties used for the *DMPK* calculation are described below.

Property 1.1. *Category of Obtainable Level of Information.*

The obtainable level of information ($LoI_{sr,pa}$) required per parameter is based on the reachable resolution provided by the data source (Table 2), adapted from the classification of [62].

- Precise measurement, enabling to reach the expected resolution and, therefore, unlocking a potential full quantification,
- Acceptable sublevel of information, enabling a semiquantitative evaluation,
- Qualitative information (e.g., yes/no; +/−; shift of tendancy (e.g., mean)),
- None.

Table 2. $LoI_{sr,pa}$—obtainable level of information.

Question: "Which Level of Information Can Be Obtained?"	
Classes	**Score**
None	0
Qualitative information	1
Acceptable sublevel of information	2
Precise measurement	3

Property 1.2. *Range Coverage Potential.*

The range coverage potential ($RaC_{sr,pa}$) concerns the completeness of a data source (i.e., the capacity for a data source to cover "all required parts of an entity's description" [88]). It can be used for characterizing a candidate dataset by answering the question "How much of the predefined range can be covered?" (Table 3).

Table 3. $RaC_{sr,pa}$—range coverage potential.

Question: "How much of the predefined range can be covered?"	
Classes	**Score**
None	0
0% to 10%	1
10% to 90%	2
90% to 100%	3

Property 1.3. *Precision Estimation.*

The precision meta-feature ($Pr_{sr,pa}$) indicates the precision of a data source, characterized through expert knowledge. The purpose is to evaluate, on the basis of experience, if the data source enables to systematically come to identical conclusions when assessing datasets acquired under repeatability conditions. This assessment is made by answering the question "Would an expert always come to the same conclusion when assessing datasets acquired under repeatability conditions?" (Table 4).

Table 4. $Pr_{sr,pa}$—precision.

Question: "Would an expert always come to the same conclusion when assessing datasets acquired under repeatability conditions?"	
Classes	**Score**
No	0
Probably to some extent	1
Yes, a priori	2

Property 1.4. *Accuracy Estimation.*

The accuracy of a data source ($Ac_{sr,pa}$) is estimated through a classification built on expert knowledge. The purpose is to assess, on the basis of experience, the potential for the acquisition method to provide measurements centered around the true value. This assessment is, thus, made by answering the question "Does the method usually enable to provide conclusions centered around the true value?" (Table 5).

Table 5. $Ac_{sr,pa}$—accuracy.

Question: Does the method usually enable to provide conclusions centered around the true value?”	
Classes	**Score**
No	0
Probably to some extent	1
Yes, a priori	2

3.3.2. Phase 2: First Degradation Factor (DF_1)—Calculation and Application

The evaluation approach of class (2) *scene* is performed by calculating the first degradation factor DF_1. To calculate DF_1, we start by applying a subcategorization of the terms reported in Table A3, Appendix B. At this stage, one mainly looks for four types of information relative to the parameters *pa* evaluated in each considered dataset *ds*:

- Where the data were acquired (acquisition area ($AA_{ds,pa}$)),
- With which spatial resolution the data were acquired (spatial resolution ($SRe_{ds,pa}$)),
- When the recording of the data was initiated ($Date_{min,ds,pa}$) and, in case several recordings of them area are available, when the recording of the data was stopped ($Date_{max,ds,pa}$) (i.e., temporal range ($TRa_{ds,pa}$)),
- With which temporal resolution the data were acquired (temporal resolution ($TRe_{ds,pa}$)).

Therefore, assuming adequately registered metadata, one can decide to only focus on the terms 2.2a “spatial” and 2.2-b “temporal” in Table A3, Appendix B, for which detailing via subcategories (acquisition area, spatial resolution, temporal range, temporal resolution) can be used to report the required information. The rest of the terms in Table A3, Appendix B are considered redundant and potentially suboptimal for a spatiotemporal characterization relevant to risk analysis.

For simplicity, we further assume that no missing information is reported regarding the terms “spatial” and “temporal”. In addition, inspired by [79] and as further detailed where required in the following, we also suggest using additional dataset quality indicators. Although those are not initially reported in the DC standard, this information can automatically be added to existing metadata. In particular, we suggest inferring new spatiotemporal related terms using additional generic data quality measures, such as the number of missing values, non-expected records, or invalid records. This choice is further discussed in Section 5.

The global DF_1 can be calculated for any dataset *ds* and related parameter *pa* as

$$DF_{1,ds,pa} = \left(1 - \|DF_{1a,ds,pa}\|\right) \times \left(1 - \|DF_{1b,ds,pa}\|\right), \tag{4}$$

where the variables are defined as follows:

- $DF_{1,ds,pa}$: first degradation factor calculated per candidate dataset *ds* and per parameter *pa*,
- $DF_{1a,ds,pa}$: first degradation factor due to spatial properties, calculated per candidate dataset *ds* and per parameter *pa*,
- $DF_{1b,ds,pa}$: first degradation factor due to temporal properties, calculated per candidate dataset *ds* and per parameter *pa*.

The calculation detailing of $DF_{1a,ds,pa}$ is presented in Section 3.3.2.1, and that of $DF_{1b,ds,pa}$ is presented in Section 3.3.2.2.

An updated score can be given to the potential of knowledge (*PK*) for any dataset *ds* and related parameter *pa* as in Equation (5).

$$PK_{DF1,ds,pa} = DMPK_{sr,pa} \times DF_{1,ds,pa}, \tag{5}$$

where the variables are defined as follows:

- $PK_{DF1,ds,pa}$: updated potential of knowledge of the dataset ds and related parameter pa after applying the first degradation factor,
- $DMPK_{sr,pa}$: default maximum potential of knowledge per data source sr and per parameter pa,
- $DF_{1,ds,pa}$: first degradation factor calculated per candidate dataset ds and per parameter pa.

The calculation of the updated potential of knowledge ($PK_{DF1,ds,pa}$) enables reconsideration and potentially reorganizing the ranking originally provided at the end of phase 1.

3.3.2.1. DF_{1a}—First Degradation Factor Due to Spatial Properties

DF_{1a}, the first element to be determined for setting up the global DF, is calculated per candidate dataset ds and per parameter pa (i.e., $DF_{1a,ds,pa}$). We consider five properties, which are further detailed below, to estimate the quality of a dataset with regard to its spatial characteristics. Those are used to determine the form of DF_{1a} through a normalized weighted sum as

$$DF_{1a,ds,pa} = \frac{x_{SC}\,\frac{SC_{ds,pa}}{3} + x_{SRe}\,\frac{SRe_{ds,pa}}{2} + x_{SDe}\,\frac{SDe_{ds,pa}}{3} + x_{SDi}\,SDi_{ds,pa} + x_{SN}\,\frac{SN_{ds,pa}}{3}}{x_{SC} + x_{SRe} + x_{SDe} + x_{SDi} + x_{SN}}, \quad (6)$$

where the variables are defined as follows:

- $DF_{1a,ds,pa}$: first degradation factor due to spatial properties, calculated per candidate dataset ds and per parameter pa,
- $SC_{ds,pa}$: spatial coverage of candidate dataset ds per parameter pa,
- $SRe_{ds,pa}$: spatial resolution of candidate dataset ds per parameter pa,
- $SDe_{ds,pa}$: spatial density of candidate dataset ds per parameter pa,
- $SDi_{ds,pa}$: spatial distribution of candidate dataset ds per parameter pa,
- $SN_{ds,pa}$: spatial noise of candidate dataset ds per parameter pa,
- x_{SC}, x_{SRe}, x_{SDe}, x_{SDi}, x_{SN}: weights given by stakeholders to the spatial coverage, spatial resolution, spatial density, spatial distribution, and spatial noise of the data, respectively.

The weights give stakeholders the possibility to manage the importance given to meta-parameters as wished. For simplicity, a naïve approach setting those weights to 1 is applied for the rest of the present paper [87].

Given a list of predefined thresholds and the score obtained for $DF_{1a,ds,pa}$, one can automatically assess whether further processing a dataset under review is meaningful; further analysis of the dataset can be postponed and only reconsidered in the absence of other relevant datasets.

Property 2.1. *Spatial Coverage* $SC_{ds,pa}$.

The spatial coverage indicates how much of the area of the selected item of interest (area of interest—*AoI*) is covered by the selected dataset (Acquisition area—*AA*). Mathematically, the percentage of spatial coverage $sc_{ds,pa}$ of a dataset ds and for a parameter pa, with the *AoI* including the item under review in the risk analysis, can be expressed as in Equation (7).

$$sc_{ds,pa} = \frac{Part\ of\ AoI\ spatially\ covered\ by\ AA_{ds,pa}}{Total\ AoI}, \quad (7)$$

where the variables are defined as follows:

- $sc_{ds,pa}$: spatial coverage of candidate dataset ds per parameter pa,
- AoI: area of interest,
- $AA_{ds,pa}$: acquisition area of candidate dataset ds and per parameter pa.

Table 6 presents the classes we propose to categorize $sc_{ds,pa}$ for assessing the meta-feature $SC_{ds,pa}$.

Table 6. $SC_{ds,pa}$—spatial coverage.

Classes	Score
Low ($sc_{ds,pa}$ < 50%)	−3
Medium (50% < $sc_{ds,pa} \leq$ 80%)	−2
High (80% < $sc_{ds,pa} \leq$ 95%)	−1
Very high ($sc_{ds,pa}$ > 95%)	0

Property 2.2. *Spatial Resolution* $SRe_{ds,pa}$.

This property is used to assess if the dataset provides the minimum required level of information in terms of spatial resolution for a specific parameter. We suggest, for a dataset *ds* and a parameter *pa*, a relative classification based on the classes reported for the estimation of the spatial extrapolation potential SEP_{pa} in the parameter characterization (i.e., NA, individual, or $<10^0$, $<10^1$, $<10^2$, $<10^3$, $\geq 10^3$). Table 7 presents the second meta-feature of learning phase 2a.

Table 7. $SRe_{ds,pa}$—spatial resolution.

Classes	Score
Distant (2 classes below or more)	−2
Close (1 class below)	−1
Sufficient (similar class or above)	0

Property 2.3. *Spatial Density* $SDe_{ds,pa}$.

This property is used to provide a statistical data quality check on the basis of the number of relevant missing values (spatially speaking). Mathematically, the spatial density $sde_{ds,pa}$ of a dataset *ds* and for a parameter *pa* can be expressed as in Equation (8).

$$sde_{ds,pa} = 1 - \frac{Number\ of\ missing\ values_{ds,pa}}{Total\ number\ of\ expected\ records_{ds,pa}}, \tag{8}$$

where $sde_{ds,pa}$ is the spatial density of candidate dataset *ds* per parameter *pa*.

The classes we propose to categorize $sde_{ds,pa}$, defining a third meta-feature of learning phase 2a, called $SDe_{ds,pa}$ (spatial density for dataset *ds* and parameter *pa*), are reported in Table 8.

Table 8. $SDe_{ds,pa}$—spatial density.

Classes	Score
Low ($sde_{ds,pa}$ < 50%)	−3
Medium (50% $\leq sde_{ds,pa}$ < 80%)	−2
High (80% $\leq sde_{ds,pa}$ < 95%)	−1
Very high ($sde_{ds,pa} \geq$ 95%)	0

Property 2.4. *Spatial Distribution* $SDi_{ds,pa}$.

This property is used to provide a statistical data quality check on the basis of the spatial distribution of missing values (spatially speaking). Mathematically, the spatial density $sdi_{ds,pa}$ of a dataset *ds* and for a parameter *pa* can be expressed as in Equation (9).

$$sdi_{ds,pa} = \frac{Average\ distance\ bewteen\ missing\ values_{ds,pa}}{Maximum\ distance\ between\ 2\ expected\ records_{ds,pa}}, \tag{9}$$

where $sdi_{ds,pa}$ is the spatial distribution of candidate dataset *ds* per parameter *pa*.

The classes we propose to categorize $sdi_{ds,pa}$, defining the fourth meta-feature of learning phase 2a, called $SDi_{ds,pa}$ (spatial distribution for dataset *ds* and parameter *pa*), are presented in Table 9.

Table 9. $SDi_{ds,pa}$—spatial distribution.

Classes	Score
Heterogeneous distribution ($0 < sdi_{ds,pa} < 30\%$)	−1
Homogeneous distribution ($sdi_{ds,pa} \geq 30\%$ or $sdi_{ds,pa} = 0$)	0

Property 2.5. *Spatial Noise* $SN_{ds,pa}$

This property is used to provide a statistical data quality check on the basis of the proportion of noise (spatially speaking). Mathematically, the spatial noise $sn_{ds,pa}$ of a dataset *ds* and for a parameter *pa* can be expressed as in Equation (10).

$$sn_{ds,pa} = \frac{Number\ of\ invalid\ values_{ds,pa}}{Total\ number\ of\ records_{ds,pa}}, \tag{10}$$

where $sn_{ds,pa}$ is the spatial noise of candidate dataset *ds* per parameter *pa*.

A fifth meta-feature of learning phase 2a, called $SN_{ds,pa}$ (spatial noise for dataset *ds* and parameter *pa*), can be used for characterizing a candidate dataset according to $sn_{ds,pa}$ (Table 10).

Table 10. $SN_{ds,pa}$—spatial noise.

Classes	Score
Low ($sn_{ds,pa} < 10\%$),	−3
Medium ($10\% \leq sn_{ds,pa} < 20\%$)	−2
High ($20\% \leq sn_{ds,pa} < 50\%$)	−1
Very high ($sn_{ds,pa} \geq 50\%$)	0

3.3.2.2. DF_{1b}—First Degradation Factor Due to Temporal Properties

DF_{1b}, the second element to be determined for setting up the global DF_1, is calculated per candidate dataset *ds* and per parameter *pa* (i.e., $DF_{1b,ds,pa}$). We consider six properties-which are further detailed below-to estimate the quality of a dataset with regard to its temporal characteristics. Those are used to determine the form of DF_{1b} through a normalized weighted sum as:

$$DF_{1b,\ ds,\ pa} = \frac{x_{TP}\ \frac{TP_{ds,\ pa}}{2} + x_{TOU}\ TOU_{ds,\ pa} + x_{TRe}\ \frac{TRe_{ds,pa}}{2} + x_{TDe}\ \frac{TDe_{ds,pa}}{3} + x_{TDi}\ TDi_{ds,pa} + x_{TN}\ \frac{TN_{ds,pa}}{3}}{x_{TP} + x_{TOU} + x_{TRe} + x_{TDe} + x_{TDi} + x_{TN}} \tag{11}$$

where the variables are defined as follows:

- $DF_{1b,ds,pa}$: first degradation factor due to temporal properties, calculated per candidate dataset *ds* and per parameter *pa*,
- $TP_{ds,pa}$: temporal pertinence of candidate dataset *ds* per parameter *pa*,
- $TOU_{ds,pa}$: temporal overlap utility of candidate dataset *ds* per parameter *pa*,
- $TRe_{ds,pa}$: temporal resolution of candidate dataset *ds* per parameter *pa*,
- $TDi_{ds,pa}$: temporal distribution of candidate dataset *ds* per parameter *pa*,
- $TN_{ds,pa}$: temporal noise of candidate dataset *ds* per parameter *pa*,
- x_{TP}, x_{TOU}, x_{TRe}, x_{TDe}, x_{TDi}, x_{TN}: weights given by stakeholders to the temporal pertinence, temporal overlap utility, temporal resolution, temporal distribution, and temporal noise of the data, respectively.

The weights give stakeholders the possibility to manage the importance given to meta-parameters as wished. For simplicity, a naïve approach setting those weights to 1 is applied for the rest of the present paper [87].

Note that the calculations of the temporal resolution $TRe_{ds,pa}$, the temporal density $TDe_{ds,pa}$, the temporal distribution $TDi_{ds,pa}$, and the temporal noise $TN_{ds,pa}$ are meaningless for datasets considered as punctual in the calculation of the temporal overlap utility $TOU_{ds,pa}$ (see details below). Therefore, those terms are not considered in the calculation of DF_{1b} in such a situation.

Property 2.6. *Temporal Pertinence* $TP_{ds,pa}$.

This property is used to assess how meaningful the exploitation of a dataset *ds* is for the analysis of a parameter *pa* in view of the age of the dataset at a given date *d* and the temporal sensitivity TS_{pa} reported in the parameter characterization (i.e., hours, days, weeks, months, years, decades, or "constant").

Mathematically, the temporal pertinence $tp_{ds,pa}$ of a dataset *ds* and for a parameter *pa* at a given date *d* can be expressed as in Equation (12).

$$tp_{ds,pa} = \frac{Date\ of\ analysis\ d - Date_{max,ds,pa}}{TS_{pa}}, \tag{12}$$

where the variables are defined as follows:

- $tp_{ds,pa}$: temporal pertinence of candidate dataset *ds* per parameter *pa*,
- TS_{pa}: temporal sensitivity of parameter *pa*,
- $Date_{max,ds,pa}$: date when the recording of the data was stopped.

The classes we propose to categorize $tp_{ds,pa}$, defining a first meta-feature of learning phase 2b, called $TP_{ds,pa}$ (temporal pertinence for dataset *ds* and parameter *pa*), are reported in Table 11.

Table 11. $TP_{ds,pa}$—temporal pertinence.

Classes	Score
Distant (2 classes below or more)	−2
Close (1 class below)	−1
Sufficient (similar class or above)	0

Property 2.7. *Temporal Overlap Utility* $TOU_{ds,pa}$.

This property enables one to qualify the utility of the temporal overlap of dataset *ds* for a parameter *pa* considering the temporal sensitivity TS_{pa} reported in the parameter characterization (i.e., hours, days, weeks, months, years, decades, or "constant"). Mathematically, the temporal overlap utility $tou_{ds,pa}$ of dataset *ds* for a parameter *pa* can be expressed as in Equation (13).

$$tou_{ds,pa} = \frac{Date_{max,ds,pa} - Date_{min,ds,pa}}{TS_{pa}}, \tag{13}$$

where the variables are defined as follows:

- $tou_{ds,pa}$: temporal overlap utility of candidate dataset *ds* per parameter *pa*,
- TS_{pa}: temporal sensitivity of parameter *pa*,
- $Date_{max,ds,pa}$: date when the recording of the data was stopped,
- $Date_{min,ds,pa}$: date when the recording of the data was initiated.

The classes we propose to categorize $tou_{ds,pa}$, defining a second meta-feature of learning phase 2b, called $TOU_{ds,pa}$ (temporal overlap utility for dataset *ds* and parameter *pa*), are reported in Table 12.

Table 12. $TOU_{ds,pa}$—temporal overlap utility.

Classes	Score
Timeseries ($tou_{ds,pa} > 1$)	−1
Punctual ($tou_{ds,pa} \leq 1$)	0

Property 2.8. *Temporal Resolution* $TRe_{ds,pa}$.

This property is used to assess if the dataset enables providing the minimum required level of information in terms of temporal resolution for a specific parameter. We suggest, for a dataset *ds* and a parameter *pa*, a relative classification based on the classes reported for the estimation of the temporal sensitivity TS_{pa} reported in the parameter characterization (i.e., hours, days, weeks, months, years, decades, or "constant"). Therefore, a third meta-feature of learning phase 2b, called $TRe_{ds,pa}$ (temporal resolution for dataset *ds* and parameter *pa*), can be used for characterizing a candidate dataset (Table 13).

Table 13. $TRe_{ds,pa}$—temporal resolution.

Classes	Score
Distant (2 classes below or more)	−2
Close (1 class below)	−1
Sufficient (similar class or above)	0

Property 2.9. *Temporal Density* $TDe_{ds,pa}$.

This property is used to provide a statistical data quality check on the basis of the number of relevant missing values (temporally speaking). Mathematically, the temporal density $tde_{ds,pa}$ of a dataset *ds* and for a parameter *pa* can be expressed as in Equation (14).

$$tde_{ds,pa} = 1 - \frac{Number\ of\ missing\ values_{ds,pa}}{Total\ number\ of\ expected\ records_{ds,pa}}, \tag{14}$$

where $tde_{ds,pa}$ is the temporal density of candidate dataset *ds* per parameter *pa*.

Table 14 presents the classes we propose to categorize $tde_{ds,pa}$, defining a fourth meta-feature of learning phase 2b, called $TDe_{ds,pa}$ (temporal density for dataset *ds* and parameter *pa*).

Table 14. $TDe_{ds,pa}$—temporal density.

Classes	Score
Low ($tde_{ds,pa} < 50\%$)	−3
Medium ($50\% \leq tde_{ds,pa} < 80\%$)	−2
High ($80\% \leq tde_{ds,pa} < 95\%$)	−1
Very high ($tde_{ds,pa} \geq 95\%$)	0

Property 2.10. *Temporal Distribution* $TDi_{ds,pa}$.

This property is used to provide a statistical data quality check on the basis of the temporal distribution of missing values (temporally speaking). Mathematically, the temporal distribution $tdi_{ds,pa}$ of a dataset *ds* and for a parameter *pa* can be expressed as in Equation (15).

$$tdi_{ds,pa} = \frac{Average\ time\ bewteen\ missing\ values_{ds,pa}}{Maximum\ time\ between\ 2\ expected\ records_{ds,pa}}, \tag{15}$$

where $tdi_{ds,pa}$ is the temporal distribution of candidate dataset *ds* per parameter *pa*.

Table 15 presents the classes we propose to categorize $tdi_{ds,pa}$, defining a fifth meta-feature of learning phase 2b, called $TDi_{ds,pa}$ (temporal distribution for dataset *ds* and parameter *pa*).

Table 15. $TDi_{ds,pa}$—temporal distribution.

Classes	Score
Heterogeneous distribution ($0 < tdi_{ds,pa} < 30\%$)	−1
Homogeneous distribution ($tdi_{ds,pa} \geq 30\%$ or $tdi_{ds,pa} = 0$)	0

Property 2.11. *Temporal Noise* $TN_{ds,pa}$.

This property is used to provide a statistical data quality check on the basis of the proportion of noise (temporally speaking). Mathematically, the temporal noise $tn_{ds,pa}$ of a dataset *ds* and for a parameter *pa* can be expressed as in Equation (16).

$$tn_{ds,pa} = \frac{Number\ of\ invalid\ values_{ds,pa}}{Total\ number\ of\ records_{ds,pa}}, \tag{16}$$

where $tn_{ds,pa}$ is the temporal noise of candidate dataset *ds* per parameter *pa*.

Table 16 presents the classes we propose for defining a sixth meta-feature of learning phase 2b, called $TN_{ds,pa}$ (temporal noise for dataset *ds* and parameter *pa*).

Table 16. $TN_{ds,pa}$—temporal noise.

Classes	Score
Very high ($tn_{ds,pa} \geq 50\%$)	−3
High ($20\% \leq tn_{ds,pa} < 50\%$)	−2
Medium ($10\% \leq tn_{ds,pa} < 20\%$)	−1
Low ($tn_{ds,pa} < 10\%$)	0

3.3.3. Phase 3: Second Degradation Factor (DF_2)—Calculation and Application

The evaluation approach of class (3) *objectives/author/circumstances* consists of calculating the second degradation factor DF_2. To calculate DF_2, we also suggest a recategorization of the terms reported in Table A4 in Appendix B on the basis of two motivations:

- We do not apply advanced natural language processing techniques in this first version of the method,
- The terms 2.9-b "modified" and 2.10-b "valid" in Table A3, Appendix B may also be used for trust assessment of a dataset.

Table 17 presents the retained terms and their associated meta-features. The management of trust-related properties consists of defining the value given to the meta-features on the basis of lists of actors, standards, references, etc. associated with predefined classes and identified over time [39]. Those meta-features are only dataset-specific and affect all parameters informed by the dataset identically.

As a result, we determine the form of the DF_2 for any dataset *ds* through a normalized weighted sum as in Equation (17).

$$\begin{aligned} DF_{2,ds} = \Big(& x_A \frac{A_{ds}}{4} + x_{BC} \frac{BC_{ds}}{2} + x_{CT} \frac{CT_{ds}}{2} + x_{Co} \frac{Co_{ds}}{2} + x_{Cr} \frac{Cr_{ds}}{2} + x_{EL} \frac{EL_{ds}}{3} + x_{HV} \frac{HV_{ds}}{2} + x_{IRefB} \frac{IRefB_{ds}}{2} + x_{IRepB}\,\mathrm{IRepB} \\ & + x_{IVO} \frac{IVO_{ds}}{2} + x_M M_{ds} + x_{Pr} \frac{Pr_{ds}}{2} + x_{Pu} \frac{Pu_{ds}}{2} + x_{Ref} \frac{\mathrm{Ref}_{ds}}{2} + x_{Rep}\,\mathrm{Rep}_{ds} + x_{Sr} \frac{Sr_{ds}}{2} + x_V V_{ds} \Big) / (x_A + x_{BC} \\ & + x_{CT} + x_{Co} + x_{Cr} + x_{EL} + x_{HV} + x_{IRefB} + x_{IRepB} + x_{IVO} + x_M + x_{Pr} + x_{Pu} + x_{Ref} + x_{Rep} + x_{Sr} + x_V \Big), \end{aligned} \tag{17}$$

where A_{ds}, BC_{ds}, CT_{ds}, Co_{ds}, Cr_{ds}, EL_{ds}, HV_{ds}, $IRefB_{ds}$, $IRepB_{ds}$, IVO_{ds}, M_{ds}, Pr_{ds}, Pu_{ds}, Ref_{ds}, Rep_{ds}, and Sr_{ds} are defined in Table 17, and x_A, x_{BC}, x_{CT}, x_{Co}, x_{Cr}, x_{EL}, x_{HV}, x_{IRefB}, x_{IRepB}, x_{IVO}, x_M, x_{Pr}, x_{Pu}, x_{Ref}, x_{Rep}, x_{Sr}, and x_V are weights given by stakeholders to the properties defined in Table 17.

Table 17. Description and categorization of trust-related meta-features.

Term	Meta-Feature	Pre-Defined Classes and Respective Values
Audience	A_{ds}	Regulatory authorities (0) Field specialists (−1) Targeted non-specialists (−2) Open access (−3) Not valued (−4)
BibliographicCitation	BC_{ds}	Official (regulations, standards, recognized journals, etc.) (0) Valued (−1) Not valued (−2)
ConformsTo	CT_{ds}	Actual (0) Depreciated (−1) Unrecognized (−2)
Contributor	Co_{ds}	Official/authorities (0) Valued (−1) Not valued (−2)
Creator	Cr_{ds}	Official/authorities (0) Valued (−1) Not valued (−2)
EducationLevel	EL_{ds}	Senior (0) Junior (−1) Trainee (−2) Not related (−3)
HasVersion	HV_{ds}	Latest version (0) Not first/not last version (−1) First version (−2)
IsReferencedBy	$IRefB_{ds}$	Official (regulations, standards, recognized journals, etc.) (0) Valued (−1) Not valued (−2)
IsReplacedBy	$IRepB_{ds}$	Nothing (0) Something (−1)
IsVersionOf	IVO_{ds}	Latest version (0) Not first/not last version (−1) First version (−2)
Modified	M_{ds}	Original file (0) Not original file (−1)
Provenance	Pr_{ds}	Official (regulations, standards, recognized journals, etc.) (0) Valued (−1) Not valued (−2)
Publisher	Pu_{ds}	Official (regulations, standards, recognized journals, etc.) (0) Valued (−1) Not valued (−2)
References	Ref_{ds}	Official (regulations, standards, recognized journals, etc.) (0) Valued (−1) Not valued (−2)
Replaces	Rep_{ds}	Something (0) Nothing (−1)
Source	Sr_{ds}	Official (regulations, standards, recognized journals, etc.) (0) Valued (−1) Not valued (−2)
Valid	V_{ds}	Valid (0) Not Valid (−1)

The weights give stakeholders the possibility to manage the importance given to meta-parameters as wished. For simplicity, a naïve approach setting those weights to 1 is applied for the rest of the present paper [87].

Following the determination of DF_2, we can update the score given to the PK for any dataset ds and related parameter pa as in Equation (18).

$$PK_{DF1,DF2,ds,pa} = PK_{DF1,ds,pa} \times (1 - \|DF_{2,ds}\|), \quad (18)$$

where the variables are defined as follows:

- $PK_{DF1,DF2,ds,pa}$: updated potential of knowledge of the dataset ds and related parameter pa after applying the first and the second degradation factors,
- $PK_{DF1,ds,pa}$: updated potential of knowledge of the dataset ds and related parameter pa after applying the first degradation factor,
- $DF_{2,ds}$: second degradation factor calculated per candidate dataset ds.

The calculation of the updated potential of knowledge ($PK_{DF1,DF2,ds,pa}$) enables a final reconsideration and potential reorganization of the dataset ranking as an output of phase 2. The result is a ranking of data sources optimized for the potential of knowledge for each of the parameters that the available datasets can inform. The application of the presently described method ensures that the data used to estimate both probabilities and consequences required for the risk analysis correspond to the best level of information available to the stakeholders, as expected by ISO 31000 [19].

4. Case Study—Power-Grid Risk Analysis

This section illustrates the method described in Section 3 through a simplified application to vegetation management of power grids. It describes the context, the hazard identification, and the application of reinforcement actions 1 and 2 (Sections 4.1 and 4.2). The three-phases method is applied in Section 4.3. The assessment is based on the evaluation of six experts specialized in risk analysis, data analytics, power-grid management, and vegetation analysis. The case study aims to illustrate the applicability and pertinence of the proposed method, rather than a full analysis covering all aspects required for executing a complete quantitative analysis. The scope is, thus, limited to large-scale power grids in Norway. Additionally, we consider only a sub-selection of parameters and a sub-selection of data sources/datasets relative to one specific dimension involved in the probability of outages due to tree fall on power lines, as detailed below.

Power grids are pillars for the good functioning of our modern and digitalized society. An important part of those networks consists of overhead power lines used for both transportation and distribution of power in regional, national, and international configurations [89]. Several hazards may compromise the integrity of those power lines. For instance, large-impact events can destroy overhead power lines, such as hurricanes, ice storms, and landslides [22]. They can also be damaged due to more local hazards, such as vegetation [5,83]. Indeed, vegetation represents a primary source of outages and has been identified as one of the root causes of some major blackouts in history [90].

Vegetation can lead to outages either via trees falling on the power lines (scenario 1) or by growing under the infrastructure until grounding one phase (scenario 2). Power-grid operators, thus, need to periodically inspect their entire network and trim vegetation in areas showing a higher probability of dangerous tree falls to avoid scenario 1. However, the way such operations are executed today (e.g., helicopter-based, foot patrols) is time-consuming, expensive, and challenging in remote and potentially hazardous areas. A risk-based approach can, thus, optimize the prioritization of actions to execute, and the decision making can be enhanced if supported by the maximum available existing data.

4.1. Reinforcement Action 1—Level of Analysis

The level of analysis for risk assessment of large-scale power grids can range from macroscopic perspectives (e.g., satellite-based inspections [20,22,91]) to a microscopic perspective (e.g., asset structural analysis [89]). Considering the nature of the infrastructure, the hazard, and resource constraints that power-grid operators usually face, we define the optimal level of analysis for risk assessment in power grids as the size of individual items (substations, power poles, etc.). We additionally break down power lines to obtain more localized items due to the extended nature of those assets. Figure 8 illustrates such a subdivision via an aerial perspective.

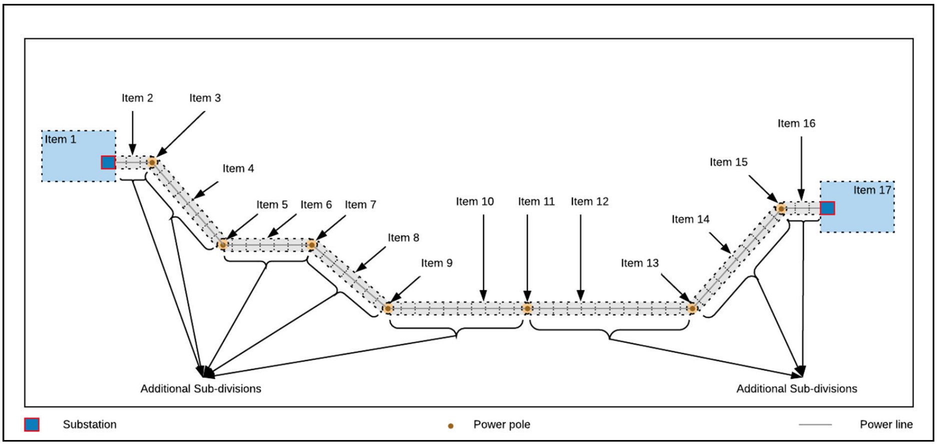


Figure 8. Section subdivision of a schematic power grid. The items of interest consist here of the substations, the power poles, and the power line subsections (Aerial 2D view).

Figure 9 shows the item of interest chosen for the risk analysis. It furthermore illustrates the presence of trees along the power lines.

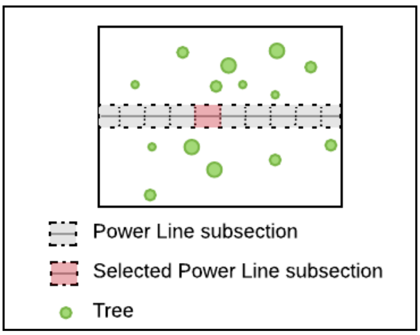


Figure 9. Selection of a power-line subsection as an item of interest for a risk analysis. The existence of the vegetation hazard is indicated by the presence of trees in the surrounding of the power line (Aerial 2D view).

4.2. Reinforcement Action 2—Parameter Characterization and Data Source Identification

Three different dimensions can be reported when estimating the probability of outage due to a tree falling on a power line.:

- The physical configuration,
- The stability of the trees surrounding the power lines,
- External factors, such as strong winds.

The following parameters play a role in the definition of the physical configuration:

- Vegetation density/number of trees (*),
- Forest social configuration (i.e., distance characterization between trees),
- Height of tree (*),
- Structure of tree crown (depth),

- Structure of tree crown (width, diameter) (*),
- Terrain exposure to wind,
- X–Y direction from a tree to the power line,
- X–Y distance from a tree to the power line (*),
- Z-delta (intensity of altitude variation).

Table 18 reports the characterization we propose for the four parameters we selected in this case study designated by (*) in the preceding list.

Table 18. Characterization of a subselection of parameters.

Parameters	Unit and Optimal Resolution	Acceptable Sub-Level of Information for Semi-Quantitative Evaluation	Minimum Range (Nominal Unit)	Spatial Extrapolation Potential (sqm): NA, Individual, or $<10^0$, $<10^1$, $<10^2$, $<10^3$, $\geq 10^3$)	Relevant Time Changing Rate (Hours, Days, Weeks, Months, Years, Decades, "Constant")
Density/number of trees	Number of trees/100 m^2	High, medium, low (e.g., percentage of tree coverage/100 m^2)	0–30	$<10^2$	Years
Height of tree	Meters (cm)	Large, medium, small	1–50	NA, individual, or $<10^0$	Months
Structure of tree crown (width, diameter)	Meters (cm)	Large, medium, small	0–30	NA, individual, or $<10^0$	Weeks
X–Y distance from a tree to the power line	Meters (cm)	High, medium, low, very low, e.g., high ($x > 30$), medium ($10 < x < 30$), low ($1 < x < 10$), very low ($x < 1$ m)	0–50	NA, individual, or $<10^0$	Weeks

Identification of data sources able to provide information for the four retained parameters is then executed. LiDAR point clouds, orthophotos based on aerial images, and satellite-based orthophotos correspond to some of the relevant data sources. The complete list of preselected sources is reported in Appendix C. Appendix C also reports suggestions of preprocessing methods usable to link each data source to the selected parameters.

4.3. Three-Phases Method Application

The geolocation of the *AoI* integrating the item of interest is given in the ETRS89/UTM32N coordinate system as follows:

- Minimum easting (X): 610,205,
- Minimum northing (Y): 6,561,098,
- Maximum easting (X): 610,253,
- Maximum northing (Y): 6,561,122.

The risk analysis is assumed to be made on 1 December 2021.

We consider a selection of three datasets to evaluate the probability of outage due to tree falls on power lines: a LiDAR point cloud, an orthophoto based on aerial images, and a satellite-based orthophoto.

The chosen files correspond to simulated realistic datasets generated for the present study. We assume having used crosswalks where required, and we report, for each dataset, the equivalent of original DC terms necessary for the analysis in Table 19. Note that the nature of the considered files and the nature of the evaluated parameters lead the spatiotemporal characteristics (acquisition area, spatial resolution, etc.) considered in the present case study to similarly impact all addressed parameters. The outcome of each phase (i.e., the ranking of the datasets based on their estimated informative potential after the application of each phase) is summarized in a unique table (Table 25) at the end of Section 4.

Table 19. DC-like terms for the three simulated datasets.

N°	Term		Dataset (a)	Dataset (b)	Dataset (c)
File-Related terms					
1-1	Format		LASF	TIFF	JPEG2000
1-2	Type		LiDAR point cloud	Orthophoto based on aerial images	Satellite-based orthophoto
Spatiotemporal-related terms					
2.2-a	Spatial				
		Acquisition area	Min. easting (X): 609,600 Min. northing (Y): 6,561,000 Max. easting (X): 610,399 Max. northing (Y): 6,561,599	Min. easting (X): 609,731 Min. northing (Y): 6,560,621 Max. easting (X): 610,639 Max. northing (Y): 6,561,425	Min. easting (X): 599,395 Min. northing (Y): 6,514,003 Max. easting (X): 638,139 Max. northing (Y): 6,601,208
		Resolution	<1 m (5 points per square meter)	0.2 m	10 m
2.3-b	Temporal				
		$Date_{min}$	25 April 2012	9 June 1989	14 June 2021
		$Date_{max}$	/	/	/
		Resolution	/	/	/
Objective/author/circumstance-related terms					
3.1	Audience		Open access		Brief standard expertise
3.4	BibliographicCitation				
3.5	ConformsTo				
3.6	Contributor				
3.7	Creator		Terratec AS	Norsk luftfoto og fjernmåling AS	ESA
3.9	EducationLevel		Not related		
3.11	HasVersion		0.0.2		
3.12	IsReferencedBy			Norge i bilder	
3.13	IsReplacedBy		NDH Østfold 5pkt 2015		
3.14	IsVersionOf				
3.15	Provenance		Kartverket Oslo	Geovekst	ESA
3.16	Publisher			Rambøll Norge AS	
3.17	References		'LAStools (c) by rapidlasso GmbH'; 'lasheight (141117) commercial'		
3.18	Replaces		None		
3.19	Source				
2.9-b	Modified		10 July 2018		
2.10-b	Valid				

4.3.1. Default Maximum Potential of Knowledge (*DMPK*)

The knowledge acquired during learning phase 1 enabling one to report the *DMPK* is detailed per data source and per parameter in Table 20.

The scores obtained via the calculation of the *DMPK* for each data source and each parameter enable generating an initial ranking per parameter of the considered datasets, as described in Table 25.

4.3.2. First Degradation Factor (DF_1)

Table 21 reports the results of calculations required for the quality assessment of (1) spatial inferred scene-related terms and (2) temporal inferred scene-related terms.

We characterize the contribution of the spatial information to the first degradation factor for each parameter informed by each of the retained dataset as reported in Table 22.

Table 20. Meta-features phase 1—*DMPK* calculation per data source per parameter.

Parameters per Source	Obtainable Level of Information $LoI_{sr,pa}$	Range Coverage Potential $RaC_{sr,pa}$	Precision Estimation $Pr_{sr,pa}$	Accuracy Potential $Ac_{sr,pa}$	Default Maximum Potential of Knowledge $DMPK_{sr,pa}$
LiDAR point cloud					
Density/number of trees	Precise measurement (3)	90% to 100% (3)	Yes, a priori (2)	Yes, a priori (2)	(3/3 + 3/3 + 2/2 + 2/2)/4 = 1
Height of tree	Precise measurement (3)	90% to 100% (3)	Yes, a priori (2)	Yes, a priori (2)	(3/3 + 3/3 + 2/2 + 2/2)/4 = 1
Structure of tree crown (width, diameter)	Precise measurement (3)	90% to 100% (3)	Yes, a priori (2)	Yes, a priori (2)	(3/3 + 3/3 + 2/2 + 2/2)/4 = 1
X–Y distance from a tree to the power line	Precise measurement (3)	90% to 100% (3)	Yes, a priori (2)	Yes, a priori (2)	(3/3 + 3/3 + 2/2 + 2/2)/4 = 1
Orthophoto based on aerial images					
Density/number of trees	Precise measurement (3)	90% to 100% (3)	Yes, a priori (2)	Probably to some extent (1)	(3/3 + 3/3 + 2/2 + 1/2)/4 = 0.875
Height of tree	Acceptable sublevel of information (2)	10% to 90% (2)	Probably to some extent (1)	Probably to some extent (1)	(2/3 + 2/3 + 1/2 + 1/2)/4 = 0.5825
Structure of tree crown (width, diameter)	Precise measurement (3)	90% to 100% (3)	Yes, a priori (2)	Yes, a priori (2)	(3/3 + 3/3 + 2/2 + 2/2)/4 = 1
X–Y distance from a tree to the power line	Precise measurement (3)	90% to 100% (3)	Probably to some extent (1)	Yes, a priori (2)	(3/3 + 3/3 + 1/2 + 2/2)/4 = 0.875
Satellite-based orthophoto					
Density/number of trees	Acceptable sublevel of information (2)	90% to 100% (3)	Probably to some extent (1)	Probably to some extent (1)	(2/3 + 3/3 + 1/2 + 1/2)/4 = 0.665
Height of tree	Acceptable sublevel of information (2)	10% to 90% (2)	No (0)	Probably to some extent (1)	(2/3 + 2/3 + 0/2 + 1/2)/4 = 0.4575
Structure of tree crown (width, diameter)	Acceptable sublevel of information (2)	10% to 90% (2)	Probably to some extent (1)	Probably to some extent (1)	(2/3 + 2/3 + 1/2 + 1/2)/4 = 0.5825
X–Y distance from a tree to the power line	Acceptable sublevel of information (2)	90% to 100% (3)	Probably to some extent (1)	Probably to some extent (1)	(2/3 + 3/3 + 1/2 + 1/2)/4 = 0.665

Table 21. Inferred scene-related DC terms.

Inferred Terms	Dataset (a)—LiDAR Point Cloud	Dataset (b)—Orthophoto Based on Aerial Images	Dataset (c)—Satellite-Based Orthophoto
Spatial inferred terms			
Spatial coverage	100%	100%	100%
Spatial resolution	Comparison per parameter of the spatial resolution $SRe_{ds,pa,}$ with the spatial extrapolation potential SEP_{pa}		
Spatial density	0.738	1	1
Spatial distribution	15%	0	0
Spatial noise	18%	0	0
Temporal inferred terms			
Temporal pertinence	Calculation per parameter of the ratio ((date of analysis $d - Date_{max,ds,pa}$)/temporal sensitivity TS_{pa})		
Temporal overlap utility	0	0	0
Temporal resolution	Not applicable (NA)	Not applicable (NA)	Not applicable (NA)
Temporal density	Not applicable (NA)	Not applicable (NA)	Not applicable (NA)
Temporal distribution	Not applicable (NA)	Not applicable (NA)	Not applicable (NA)
Temporal noise	Not applicable (NA)	Not applicable (NA)	Not applicable (NA)

The spatial overlap's nature justifies further proceeding with a temporal analysis of the degradation factor DF_1.

We characterize the contribution of the temporal information to the first degradation factor for each parameter informed by each of the retained dataset as reported in Table 23. Note that the detailing of the temporal resolution $TRe_{ds,pa}$, the temporal density $TDe_{ds,pa}$, the temporal distribution $TDi_{ds,pa}$, and the temporal noise $TN_{ds,pa}$ is not reported here because they were identified as not applicable on the basis of the calculation of temporal overlap utility $TOU_{ds,pa}$.

Table 22. Meta-features phase 2—spatial contribution to DF_1 per data source per parameter.

Parameters per Data Source	Spatial Coverage $SC_{ds,pa}$	Spatial Resolution $SRe_{ds,pa}$	Spatial Density $SDe_{ds,pa}$	Spatial Distribution $SDi_{ds,pa}$	Spatial Noise $SN_{ds,pa}$	DF_{1a} Spatial contribution to DF_1
LiDAR point cloud						
Density/number of trees	Very high (0)	Sufficient (0)	Medium (−2)	Heterogeneous distribution (−1)	Medium (−1)	(0/3 + 0/2 + (−2)/3 + (−1) + (−1)/3)/5 = −0.4
Height of tree	Very high (0)	Close (−1)	Medium (−2)	Heterogeneous distribution (−1)	Medium (−1)	(0/3 + (−1)/2 + (−2)/3 + (−1) + (−1)/3)/5 = −0.5
Structure of tree crown (width, diameter)	Very high (0)	Close (−1)	Medium (−2)	Heterogeneous distribution (−1)	Medium (−1)	(0/3 + (−1)/2 + (−2)/3 + (−1) + (−1)/3)/5 = −0.5
X–Y distance from a tree to the power line	Very high (0)	Close (−1)	Medium (−2)	Heterogeneous distribution (−1)	Medium (−1)	(0/3 + (−1)/2 + (−2)/3 + (−1) + (−1)/3)/5 = −0.5
Orthophoto based on aerial images						
Density/number of trees	Very high (0)	Sufficient (0)	Very high (0)	Homogeneous distribution (0)	Low (0)	(0/3 + 0/2 + 0/3 + 0 + 0/3)/5 = 0
Height of tree	Very high (0)	Sufficient (0)	Very high (0)	Homogeneous distribution (0)	Low (0)	(0/3 + 0/2 + 0/3 + 0 + 0/3)/5 = 0
Structure of tree crown (width, diameter)	Very high (0)	Sufficient (0)	Very high (0)	Homogeneous distribution (0)	Low (0)	(0/3 + 0/2 + 0/3 + 0 + 0/3)/5 = 0
X–Y distance from a tree to the power line	Very high (0)	Sufficient (0)	Very high (0)	Homogeneous distribution (0)	Low (0)	(0/3 + 0/2 + 0/3 + 0 + 0/3)/5 = 0
Satellite-based orthophoto						
Density/number of trees	Very high (0)	Sufficient (0)	Very high (0)	Homogeneous distribution (0)	Low (0)	(0/3 + 0/2 + 0/3 + 0 + 0/3)/5 = 0
Height of tree	Very high (0)	Close (−1)	Very high (0)	Homogeneous distribution (0)	Low (0)	(0/3 + (-1)/2 + 0/3 + 0 + 0/3)/5 = −0.1
Structure of tree crown (width, diameter)	Very high (0)	Close (−1)	Very high (0)	Homogeneous distribution (0)	Low (0)	(0/3 + (-1)/2 + 0/3 + 0 + 0/3)/5 = −0.1
X–Y distance from a tree to the power line	Very high (0)	Close (−1)	Very high (0)	Homogeneous distribution (0)	Low (0)	(0/3 + (-1)/2 + 0/3 + 0 + 0/3)/5 = −0.1

Table 23. Meta-features phase 2—temporal contribution to DF_1 per dataset per parameter.

Parameters per Dataset	Temporal Pertinence $TP_{ds,pa}$	Temporal Overlap Utility $TOU_{ds,pa}$	DF_{1b} Temporal Contribution to DF_1
LiDAR point cloud			
Density/number of trees	Sufficient (0)	Punctual (−1)	(0/2 + (−1))/2 = −0.5
Height of tree	Close (−1)	Punctual (−1)	((−1)/2 + (−1))/2 = −0.75
Structure of tree crown (width, diameter)	Distant (−2)	Punctual (−1)	((−2)/2 + (−1))/2 = −1
X–Y distance from a tree to the power line	Distant (−2)	Punctual (−1)	((−2)/2 + (−1))/2 = −1
Orthophoto based on aerial images			
Density/number of trees	Close (−1)	Punctual (−1)	((−1)/2 + (−1))/2 = −0.75
Height of tree	Distant (−2)	Punctual (−1)	((−2)/2 + (−1))/2 = −1
Structure of tree crown (width, diameter)	Distant (−2)	Punctual (−1)	((−2)/2 + (−1))/2 = −1
X–Y distance from a tree to the power line	Distant (−2)	Punctual (−1)	((−2)/2 + (−1))/2 = −1
Satellite-based orthophoto			
Density/number of trees	Sufficient (0)	Punctual (−1)	(0/2 + (−1))/2 = −0.5
Height of tree	Sufficient (0)	Punctual (−1)	(0/2 + (−1))/2 = −0.5
Structure of tree crown (width, diameter)	Close (−1)	Punctual (−1)	((−1)/2 + (−1))/2 = −0.75
X–Y distance from a tree to the power line	Close (−1)	Punctual (−1)	((−1)/2 + (−1))/2 = −0.75

We combine DF_{1a} and DF_{1b} to calculate the total degradation factor DF_1. This enables calculating the updated potential of knowledge ($PK_{DF1,ds,pa}$) per dataset and per parameter, as described in Table 25.

4.3.3. Second Degradation Factor (DF_2)

Metadata relative to the objectives, the authors, and the circumstances of the data acquisition enable calculating the value of the second degradation factor for each of the datasets, as detailed in Table 24.

Table 24. Meta-features phase 3—DF_2 per dataset.

Terms	Meta-Feature	Dataset (a)—LiDAR Point Cloud	Dataset (b)—Orthophoto Based on Aerial Images	Dataset (c)—Satellite-Based Orthophoto
Audience	A_{ds}	Open access (−3)	-	Field specialists (−1)
BibliographicCitation	BC_{ds}	-	-	-
ConformsTo	CT_{ds}	-	-	-
Contributor	Co_{ds}	-	-	-
Creator	Cr_{ds}	Valued (−1)	Valued (−1)	Official/authorities (0)
EducationLevel	EL_{ds}	Not related (−3)	-	-
HasVersion	HV_{ds}	First version (−2)	-	-
IsReferencedBy	$IRefB_{ds}$	-	Official (regulations, standards, recognized journals) (0)	-
IsReplacedBy	$IRepB_{ds}$	Something (−1)	-	-
IsVersionOf	IVO_{ds}	-	-	-
Provenance	Pr_{ds}	Official (regulations, standards, recognized journals) (0)	Official (regulations, standards, recognized journals) (0)	Official (regulations, standards, recognized journals) (0)
Publisher	Pu_{ds}	-	Valued (−1)	-
References	Ref_{ds}	Valued (−1)	-	-
Replaces	Rep_{ds}	Nothing (−1)	-	-
Source	Sr_{ds}	-	-	-
Modified	M_{ds}	Not original file (−1)	-	-
Valid	V_{ds}	-	-	-
$DF_{2,\,ds} = (\frac{A_{ds}}{4} + \frac{BC_{ds}}{2} + \frac{CT_{ds}}{2} + \frac{Co_{ds}}{2} + \frac{Cr_{ds}}{2} + \frac{EL_{ds}}{3} + \frac{HV_{ds}}{2} + \frac{IRefB_{ds}}{2} + IRepB_{ds} + \frac{IVO_{ds}}{2} + \frac{Pr_{ds}}{2} + \frac{Pu_{ds}}{2} + \frac{Ref_{ds}}{2} + Rep_{ds} + \frac{Sr_{ds}}{2} + M_{ds} + V_{ds})/17$		((−3)/4 + (0)/2 + (0)/2 + (0)/2 + (−1)/2 + (−3)/3 + (−2)/2 + (0)/2 + (1) + (0)/2 + (0)/2 + (0)/2 + (−1)/2 + (−1) + (0)/2 + (−1) + (0))/17 = −0.397	((0)/4 + (0)/2 + (0)/2 + (0)/2 + (−1)/2 + (0)/3 + (0)/2 + (0)/2 + (0) + (0)/2 + (0)/2 + (−1)/2 + (0)/2 + (0) + (0)/2 + (0) + (−0))/17 = −0.059	((−1)/4 + (0)/2 + (0)/2 + (0)/2 + (0)/2 + (0)/3 + (0)/2 + (0)/2 + (0) + (0)/2 + (0)/2 + (0)/2 + (0)/2 + (0) + (0)/2 + (0) + (0))/17 = −0.015

Note: Note that an optimistic approach was applied in the absence of metadata, leading all the meta-features to be set to 0 in the absence of information. An alternative pessimistic approach penalizing the absence of information could also have been applied, putting all the scores to the most degraded level. This choice and the effects of this choice on the results are further discussed in Section 5.

By applying the degradation factor DF_2 and updating the calculation of the potential of knowledge ($PK_{DF1,DF2,ds,pa}$) per dataset and per parameter, we obtain a final ranking of the best candidate dataset for each of the considered parameter, as described in Table 25.

Table 25 details the evolution of the ranking of the best candidate for each parameter based on the application of the 3 phases mentioned in our method.

The coloring (green, yellow, orange) of the scores within each phase as detailed in Table 25 reports the first, second, and third best candidate datasets for the characterization of each parameter. Table 25 indicates that LiDAR-based point clouds initially constitute the most interesting type of dataset for the present case study. Therefore, they should be preferred by default, in the absence of any other type of information relative to file properties.

However, the application of the degradation factors DF_1 and DF_2 leads to a new ranking of the best candidate datasets for each considered parameter. In particular, the satellite-based orthophoto obtains the first position in the ranking for all parameters after the application of DF_1 and keeps this position after application of DF_2. The LiDAR point cloud obtains the second position for informing the parameter "height of tree" after the application of DF_1 and keeps this position after application of DF_2. The orthophoto based on aerial images obtains the third position for the characterization of the parameter "density/number of trees" after the application of DF_1, but is reranked second after application of DF_2.

Table 25. Spatiotemporal- and trust-influenced informative potential assessment per datasets per parameter.

	Phase 1	Phase 2		Phase 3	
Parameters per Data Source	$DMPK_{sr,pa}$	$DF_{1,ds,pa}$	$PK_{DF1,ds,pa}$	$1 - \lVert DF_{2,ds} \rVert$	$PK_{DF1,DF2,ds,pa}$
LiDAR point cloud					
Density/number of trees	1	0.3	0.3	0.6029	0.1809
Height of tree	1	0.125	0.125	0.6029	0.0754
Structure of tree crown (width, diameter)	1	0	0	0.6029	0
X–Y distance from a tree to the power line	1	0	0	0.6029	0
Orthophoto based on aerial images					
Density/number of trees	0.875	0.25	0.2188	0.941	0.2058
Height of tree	0.5825	0	0	0.941	0
Structure of tree crown (width, diameter)	1	0	0	0.941	0
X–Y distance from a tree to the power line	0.875	0	0	0.941	0
Satellite-based orthophoto					
Density/number of trees	0.665	0.5	0.3325	0.985	0.3275
Height of tree	0.4575	0.45	0.2059	0.985	0.2028
Structure of tree crown (width, diameter)	0.5825	0.225	0.1311	0.985	0.1291
X–Y distance from a tree to the power line	0.665	0.225	0.1496	0.985	0.1474

5. Discussion

5.1. Method Benefits and Contribution for Risk Assessment

The accessibility to many data sources for risk assessment is a recent phenomenon for many application areas. While this consists of a great opportunity for data-driven assessments and DRA, the process for choosing one dataset to inform a parameter in detriment of another was not yet formalized. The method proposed in this paper formalizes this process. One of the main advantages of the method is the use of metadata; the method provides the ranking of the best datasets according to their informative potential without the need for the analyst to open the file and assess its content. In addition to identifying the best dataset for each parameter, the proposed approach also identifies the best parameters per source. Such information can be used as an additional indicator for strategical decision making when deciding about investments related to future data acquisitions.

The results of the method application are, to the best of our knowledge, representative of the reality that power-grid operators can face. The impact of the degradation factors highlights that a dataset originating from a less valued data source may be preferred to a dataset originating from an initially higher-ranked data source, the content of which is outdated or of too low quality. The results also show that, while getting degraded and losing its leading position to inform a specific parameter, a dataset may still remain the best choice for another parameter. This indicates the importance of assessing the informative potential of a dataset on a parameter level rather than as one block.

Moreover, the method provides flexibility for the analyst to achieve the following:

- Tune the meta-features used to calculate *DMPK* in phase 1, if expertise/follow-up gain of knowledge shows that the initial estimation was not adequate, the initial estimation needed to be updated, or if the evolution in technologies/competencies of stakeholders enables improving the initially obtainable quality of information;
- Adequately maintain elements required for the calculation of the trust-related meta-features by adding, confirming, or removing entities in the lists used for the calculation of the second degradation factor (e.g., new standard or withdraw of a previous standard);
- Have the process iterated over time (even without new datasets) and readapt the ranking of the considered sources if required;
- Modify the weights given to any of the meta-features proposed in phases 1, 2, and 3 on the basis of what one decides to be important or if new risk evidence implies that changes are required;

- Assess the potential of new types of data sources not yet known and integrate the related datasets into the risk analysis by running them through the three phases.

This approach, thus, supports the development and implementation of DRAs by ensuring proper and dynamic dataset selection in an environment with ever-increasing access to more information. It should be noted that the main benefits of the method are to be gained in a long-term application, i.e., when used after the first time, the analyst needs to update it only in case new data sources are considered.

5.2. Limitations and Further Requirements

5.2.1. Reliance on Metadata Format

The proposed method is particularly sensitive to metadata existence and quality. Yet, experience shows that metadata can be in the wrong format (i.e., not respecting existing standards), inexistent, or incomplete, as illustrated by the absence of some *objective/author/circumstance*-related terms in the case study. Metadata may also be corrupted and contain inaccurate information, due to human actions or inadequate automatic processing. Metadata in the wrong format may lead to more challenging and time-consuming conversions into DC terms, as one would have to work with/generate nonconventional crosswalks. As a partial solution, and assuming that the number of metadata terms is expected to remain limited, one may create lists of "standard-like" terms to convert non-standardized metadata into information usable for the method. In line with this concept, one could also generate further quality indicators of the observed datasets by implementing and counting alarms used to highlight missing metadata. Additionally, it might be strategic to enable one to assign some "privileges" to specific datasets to avoid naïvely degrading the value attributed to datasets due to absence of trust information in the metadata (for instance, for data originating from reference entities such as governmental-based institutions). Such approaches may also be considered when acknowledging that some datasets initially lack metadata but are known to have been generated internally and can, thus, be considered as more "trustable".

Lastly, two additional approaches—optimistic or pessimistic—may be chosen to face incomplete metadata reporting, as applied in the presented case study. For the former, stakeholders can apply the minimum degradation and generate alerts for manual verification of the final ranking when missing information has been detected. Alternatively, stakeholders may be more conservative and apply the maximum degradation to minimize the possibilities of building on uncertainties. An optimistic approach was chosen for the calculation of the second degradation factor in our work. The pessimistic approach would have led to the same ranking in the present case study because of the metadata we decided to report for those simulated datasets. However, this may usually not be the case. The choice of the strategy to follow is a more subjective task that, thus, needs to be addressed on a case-by-case analysis and is hardly generalizable in the generic description of our approach.

5.2.2. Three-Phases Method Elements

The characterization of the properties in phases 1, 2, and 3 of the method mainly consists of the definition of classes and class boundaries relative to each property, according to the authors' experience. Yet, the method provides sufficient flexibility for the boundaries to be adapted if needed, especially in a context other than power-grid management. The choices made in the definition of the three-phases method may, thus, be seen as an illustration of a general guideline that can be adapted to the context in which it is applied rather than as a rigid formulation.

The weights associated with the meta-features of each phase are likely to be different from application to application or from organization to organization. Although the approach is already implementable as is for any project where the requirements related to the learning phases are fulfilled, another implementation would require first defining the value of the reported weights. Different approaches may be considered for this purpose, and one may, for instance, tackle this problem as a meta-learning task. Alternatively, one could use more straightforward solutions such as the application of Zip's law, as applied in other work [92].

We chose a normalized weighted sum to summarize the results of all the properties used for the definition of the *DMPK* and *DFs*. This choice is considered robust, well-known, and straightforward [93,94], and it was adopted for a first formulation of the proposed methodology. Multicriteria decision-making methods could also be considered, for instance, if further dependencies between properties were to be considered in future applications of the method. Furthermore, the approach considers, right now, all datasets to be independent and analyzed independently and not leveraging one another. Future extensions should address the existence of links between datasets.

The implementation of the method is more labor-intensive in the first iterations of the process, especially when the knowledge of the involved stakeholders needs to be converted into information exploitable for the use of the presented method. Verifications required after detecting new terms in the lists used to calculate the second degradation factor may be particularly time-consuming. However, this workload and the general need for manual verifications are expected to diminish over time as the number of processed datasets increases, facilitating future automatic processing.

6. Conclusions

Despite being synonymous with considerable advantages, increasing data availability related to energy systems also implies numerous challenges from a data management perspective. In the present work, we suggest an approach enabling to tackle such challenges in the risk analysis field by analyzing features that may degrade the information potential of a dataset compared to a baseline (default maximum potential of knowledge (*DMPK*)). The key contributions of this research are twofold. First, it proposes a method for reinforcing data-related risk analysis steps. The use of this method ensures that risk analysts can methodically identify and assess the available data for informing the risk analysis key parameters. Second, it develops a method for selecting the best datasets according to their informative potential. The method, thus, formalizes, in a traceable and reproducible manner, the process for choosing one dataset to inform a parameter in detriment of another, which can lead to more accurate risk analyses.

The application of the method to vegetation-related risk assessment in power grids shows that the approach enables dynamically selecting the best information possible. It also demonstrates that a dataset originating from an initially less valued data source (the satellite-based orthophoto) may be preferred to a dataset originating from a higher-ranked data source, the content of which is outdated or of too low quality (LiDAR-based point clouds).

The proposed method reduces approximations and uncertainties in risk analysis by improving tracing of information and measurement characterization, thereby supporting the optimization of dataset preparation and integration for more efficient risk assessments. The generic aspect of the approach opens up further applications, such as management of safety barriers or risk management applied in other areas such as finance or insurance, which can also access large data lakes. The method is particularly valuable for the risk assessment of large-scale "open" systems, subject to environmental changes and presenting increasing opportunities for data capture. Such systems include not only power grids, but also transportation systems and gas pipelines. In addition to pressure, temperature, and flow sensors that might identify a failure after a leak is already established, external pipelines can benefit from data captured by autonomous systems (drones) or satellites for identifying a growing crack that can potentially result in a leak.

Lastly, although the exploitation of natural language processing techniques may in the future further support the application of the method (i.e., via advanced text-based quality assessment of metadata), there is an undiscussable need for proper and better metadata registration and maintenance. The community should encourage good practices in this direction, and our work can be used to illustrate the potential benefits of doing so.

Author Contributions: Conceptualization, M.F.P., M.R., D.R. and N.P.; methodology, M.F.P., M.R. and N.P.; validation, M.F.P., M.R., D.R., C.T.E. and N.P.; formal analysis, M.F.P.; investigation, M.F.P.; data curation, M.F.P.; writing—original draft preparation, M.F.P.; writing—review and editing, M.R., D.R., C.T.E. and N.P.; supervision, D.R. and N.P.; project administration, N.P.; funding acquisition, D.R. and N.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work is part of the project "Dynamic risk management for smart grids in large-scale interconnected power systems", funded by eSmart Systems and the Norwegian Research Council (Nærings PhD program, Project No.: 276404), which the authors would like to thank for their support.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1 reports an extract of the "open directory of metadata standards applicable to scientific data" [35], provided by the Metadata Standards Directory Working Group [33], a working group from the Research Data Alliance [34]. The directory is community-maintained.

Table A1. Extract of commonly used metadata standards [35].

Name	Original Field of Application	Description	Link to Resource
MARC	Arts and humanities	MARC (machine-readable cataloging) is a standard mostly originally used for reporting and exchanging bibliographic records.	http://www.loc.gov/marc/ (accessed on 15 February 2022)
Darwin Core	Life sciences	A body of standards used for reporting and exchanging biology-related information.	http://rs.tdwg.org/dwc/index.htm (accessed on 15 February 2022)
EML	Life sciences	EML (ecological metadata language) is a metadata specification used for reporting and exchanging ecology-related information	https://nceas.github.io/eml/ (accessed on 15 February 2022)
ISO 19115	Physical sciences and mathematics	ISO 19115 (geographic information—metadata) is schema used for the description of geographic information and services.	https://www.iso.org/standard/26020.html (accessed on 15 February 2022) https://www.iso.org/standard/53798.html (accessed on 15 February 2022)
Dublin Core	General research data	Authoritative specification of all metadata terms maintained by the Dublin Core™ Metadata Initiative (DCMI). These terms are intended to be used in combination with metadata terms from other, compatible vocabularies.	http://dublincore.org (accessed on 15 February 2022)

Appendix B

Tables A2–A4 detail the recategorization of the DC terms relevant for a risk analysis into three classes: (1) *file* (nature of the dataset), (2) *scene* (site-/time-specifications of the dataset), and (3) *objectives/author/circumstances* (agents and factors influencing data management). The tables also report the respective definition of each of the selected terms [83] (The reader is directed to the home page of the Dublin Core™ Metadata Initiative for a complete description of the terms not used in the present paper.).

Table A2. *File*-related DC terms.

N°	Term	Definition
1-1	Format	The file format, physical medium, or dimensions of the resource.
1-2	Type	The nature or genre of the resource.

Table A3. *Scene*-related DC terms.

N°	Term	Definition
2.1	Coverage	The spatial or temporal topic of the resource, spatial applicability of the resource, or jurisdiction under which the resource is relevant.
2.2-a	Spatial	Spatial characteristics of the resource (sub-property of coverage).
2.3-b	Temporal	Temporal characteristics of the resource (sub-property of coverage).
2.4-b	Date	A point or period of time associated with an event in the lifecycle of the resource.
2.5-b	Created	Date of creation of the resource (sub-property of date).
2.6-b	Issued	Date of formal issuance of the resource (sub-property of date).
2.7-b	AccrualPeriodicity	The frequency with which items are added to a collection.
2.8-b	Available	Date that the resource became or will become available (sub-property of date).
2.9-b	Modified	Date on which the resource was changed (sub-property of date).
2.10-b	Valid	Date (often a range) of validity of a resource (sub-property of date).
2.11-b	DateCopyrighted	Date of copyright of the resource (sub-property of date).
2.12-b	DateSubmitted	Date of submission of the resource (sub-property of date).
2.13-b	DateAccepted	Date of acceptance of the resource (sub-property of date).

Table A4. Objective/author/circumstance-related DC-terms.

N°	Term	Definition
3.1	Audience	A class of agents for whom the resource is intended or useful.
3.2	Abstract	A summary of the resource.
3.3	AccrualMethod	The method by which items are added to a collection.
3.4	BibliographicCitation	A bibliographic reference for the resource.
3.5	ConformsTo	An established standard to which the described resource conforms.
3.6	Contributor	An entity responsible for making contributions to the resource.
3.7	Creator	An entity responsible for making the resource.
3.8	Description	An account of the resource.
3.9	EducationLevel	A class of agents, defined in terms of progression through an educational or training context, for which the described resource is intended.
3.10	Extent	The size or duration of the resource.
3.11	HasVersion	A related resource that is a version, edition, or adaptation of the described resource.
3.12	IsReferencedBy	A related resource that references, cites, or otherwise points to the described resource.
3.13	IsReplacedBy	A related resource that supplants, displaces, or supersedes the described resource.
3.14	IsVersionOf	A related resource of which the described resource is a version, edition, or adaptation.
3.15	Provenance	A statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation.
3.16	Publisher	An entity responsible for making the resource available.
3.17	References	A related resource that is referenced, cited, or otherwise pointed to by the described resource.
3.18	Replaces	A related resource that is supplanted, displaced, or superseded by the described resource.
3.19	Source	A related resource from which the described resource is derived.
3.20	Subject	A topic of the resource.
3.21	Title	A name given to the resource.

Appendix C

Table A5 lists all data sources identified as able to provide information for the four parameters retained in the case study. The table also reports suggestions of preprocessing methods usable to link each data source to the selected parameters.

Identical cells in the table are merged where possible for consistency purposes. The value of the information that can be provided by each of the retained sources is detailed in the application of the method presented in Section 4.

Table A5. Suggestions of preprocessing methods usable to link the data sources to the selected parameters.

Data Source	Density/Number of Trees	Height of Tree	Structure of Tree Crown (Width, Diameter)	X–Y Distance from a Tree to the Power Line
Aerial optical inspection images	Visual estimation, counting	Visual estimation, classification		
Forest survey (map)	Average value reported over a pixel		-	
LiDAR point clouds	Cloud segmentation and counting or point cloud density calculation	Cloud segmentation and measurement		
Meshed photogrammetry-based point clouds	Evaluation of number, depth, and relative proportion of valleys	Mesh segmentation and measurement		
Orthophotos (aerial images)	Counting, counting per area	Visual estimation, extrapolated from crown width	Crown size measurement	Distance measurement tree-power line
Orthophotos (satellite images)	Counting, counting per area	Visual estimation, extrapolated from crown width	Crown size estimation	Distance estimation tree, power line
Pests/fungi survey (map)	Probabilistic estimation based on pests/fungi-related damages over time			-
Photogrammetry point clouds	Cloud segmentation and counting or point cloud density calculation	Cloud segmentation and measurement		
Soil survey (map)	Probabilistic estimation of having a tree based on soil type	Probabilistic estimation of having a tree and estimation of growth potential for trees depending on soil type		-
TOPEX (topographical wind exposure) (map)	Probabilistic estimation of having a tree and estimation of growth potential for trees depending on altitude + probable wind impact over time			Probabilistic estimation of having a tree and estimation of growth potential for trees depending on altitude + probable wind impact over time + Z-delta measurement for difference due to terrain variations
Topography (map)	Probabilistic estimation of having a tree and estimation of growth potential for trees depending on altitude			Probabilistic estimation of having a tree and estimation of growth potential for trees depending on altitude + Z-delta measurement for difference due to terrain variations
Weather historical data	Probabilistic estimation of having a tree and estimation of growth potential for trees depending on weather conditions			-

References

1. Zúñiga, A.A.; Baleia, A.; Fernandes, J.; Branco, P.J.D.C. Classical Failure Modes and Effects Analysis in the Context of Smart Grid Cyber-Physical Systems. *Energies* **2020**, *13*, 1215. [CrossRef]
2. Zhang, Z.; Yang, H.; Yin, X.; Han, J.; Wang, Y.; Chen, G. A Load-Shedding Model Based on Sensitivity Analysis in on-Line Power System Operation Risk Assessment. *Energies* **2018**, *11*, 727. [CrossRef]
3. Kumar, G.V.B.; Sarojini, R.K.; Palanisamy, K.; Padmanaban, S.; Holm-Nielsen, J.B. Large Scale Renewable Energy Integration: Issues and Solutions. *Energies* **2019**, *12*, 1996. [CrossRef]
4. Pacevicius, M.; Roverso, D.; Rossi, P.S.; Paltrinieri, N. Smart grids: Challenges of processing heterogeneous data for risk assessment. In Proceedings of the 14th International Conference on Probabilistic Safety Assessment and Management, Los Angeles, CA, USA, 16–21 September 2018.
5. Pacevicius, M.; Ramos, M.A.; Paltrinieri, N. Optimizing Technology-based Decision-support for management of Infrastructures under risk: The Case of Power Grids. In Proceedings of the 30th ESREL-15th PSAM, Venice, Italy, 1–5 November 2020; Research Publishing: Singapore, 2020; p. 8. [CrossRef]
6. Ciapessoni, E.; Cirio, D.; Pitto, A.; Marcacci, P.; Lacavalla, M.; Massucco, S.; Silvestro, F.; Sforna, M. A Risk-Based Methodology and Tool Combining Threat Analysis and Power System Security Assessment. *Energies* **2018**, *11*, 83. [CrossRef]
7. Sayed, A.; El-Shimy, M.; El-Metwally, M.; Elshahed, M. Reliability, Availability and Maintainability Analysis for Grid-Connected Solar Photovoltaic Systems. *Energies* **2019**, *12*, 1213. [CrossRef]
8. Elsisi, M.; Tran, M.-Q.; Mahmoud, K.; Lehtonen, M.; Darwish, M.M.F. Deep Learning-Based Industry 4.0 and Internet of Things towards Effective Energy Management for Smart Buildings. *Sensors* **2021**, *21*, 1038. [CrossRef] [PubMed]
9. Tran, M.-Q.; Elsisi, M.; Mahmoud, K.; Liu, M.-K.; Lehtonen, M.; Darwish, M.M.F. Experimental Setup for Online Fault Diagnosis of Induction Machines via Promising IoT and Machine Learning: Towards Industry 4.0 Empowerment. *IEEE Access* **2021**, *9*, 115429–115441. [CrossRef]
10. Elsisi, M.; Tran, M.; Mahmoud, K.; Mansour, D.A.; Lehtonen, M.; Darwish, M.M. Towards Secured Online Monitoring for Digitalized GIS Against Cyber-Attacks Based on IoT and Machine Learning. *IEEE Access* **2021**, *9*, 78415–78427. [CrossRef]
11. Zhang, D.; Qian, L.; Mao, B.; Huang, C.; Huang, B.; Si, Y. A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and Xgboost. *IEEE Access* **2018**, *6*, 21020–21031. [CrossRef]
12. Beerens, H.; Post, J.; Uijtdehaag, P. The use of generic failure frequencies in QRA: The quality and use of failure frequencies and how to bring them up-to-date. *J. Hazard. Mater.* **2006**, *130*, 265–270. [CrossRef]
13. Villa, V.; Paltrinieri, N.; Khan, F.; Cozzani, V. Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. *Saf. Sci.* **2016**, *89*, 77–93. [CrossRef]
14. Paltrinieri, N.; Khan, F.; Cozzani, V. Coupling of advanced techniques for dynamic risk management. *J. Risk Res.* **2015**, *18*, 910–930. [CrossRef]
15. Paltrinieri, N.; Khan, F.I. Dynamic risk analysis—Fundamentals. In *Advanced Methods of Risk Assessment and Management*; Khan, F.I., Amyotte, P.R., Eds.; Elsevier: Amsterdam, The Netherlands, 2020; Volume 4, pp. 35–60.
16. Pacevicius, M.; Paltrinieri, N.; Thieme, C.A.; Rossi, P.S. Addressing the Importance of Data Veracity during Data Acquisition for Risk Assessment Processes. In Proceedings of the 67th Annual Reliability and Maintainability Symposium, Orlando, FL, USA, 24–27 May 2021; p. 7. [CrossRef]
17. Xing, J.; Zeng, Z.; Zio, E. A framework for dynamic risk assessment with condition monitoring data and inspection data. *Reliab. Eng. Syst. Saf.* **2019**, *191*, 106552. [CrossRef]
18. Chen, N.; Liu, W.; Bai, R.; Chen, A. Application of computational intelligence technologies in emergency management: A literature review. *Artif. Intell. Rev.* **2019**, *52*, 2131–2168. [CrossRef]
19. *ISO 31000:2018*; Risk Management—Guidelines. International Standardization Organization (ISO): Geneva, Switzerland, 2018; p. 34.
20. Gazzea, M.; Pacevicius, M.; Dammann, D.O.; Sapronova, A.; Lunde, T.M.; Arghandeh, R. Automated Power Lines Vegetation Monitoring using High-Resolution Satellite Imagery. *Trans. Power Deliv.* **2021**, *37*, 308–316. [CrossRef]
21. Pacevicius, M.; Gazzea, M.; Dammann, D.O.; Sapronova, A.; Lunde, T.M.; Arghandeh, R. *GridEyeS Deliverable D3: Technical Feasibility Assessment*; StormGeo: Bergen, Norway, 2020.
22. Pacevicius, M.; Haskins, C.; Paltrinieri, N. Supporting the Application of Dynamic Risk Analysis to Real-World Situations using Systems Engineering: A focus on the Norwegian Power Grid Management. In *Recent Trends and Advances in Model Based Systems Engineering*; Springer: Cham, Switerland, 2022. [CrossRef]
23. Merriam-Webster Metadata. Merriam-Webster.com Dictionary. 2022. Available online: https://www.merriam-webster.com/dictionary/metadata (accessed on 15 February 2022).
24. Wierling, A.; Schwanitz, V.J.; Altinci, S.; Bałazińska, M.; Barber, M.J.; Biresselioglu, M.E.; Burger-Scheidlin, C.; Celino, M.; Demir, M.H.; Dennis, R.; et al. FAIR Metadata Standards for Low Carbon Energy Research—A Review of Practices and How to Advance. *Energies* **2021**, *14*, 6692. [CrossRef]
25. Song, G.; Nie, Y.; Chen, G.; Liu, X. Applied research of data-driven multi-level reliability model in transportation safety risk evaluation. In Proceedings of the 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Fuzhou, China, 12–14 June 2020; pp. 116–121.
26. *ISO/IEC 11179-3:2013*; Information Technology—Metadata Registries (MDR)—Part 3: Registry Metamodel and Basic Attributes. International Standardization Organization (ISO): Geneva, Switzerland, 2013; p. 227.

27. Wang, H.; Ye, Z. An ETL services framework based on metadata. In Proceedings of the 2010 2nd International Workshop on Intelligent Systems and Applications, Wuhan, China, 22–23 May 2010. [CrossRef]
28. Bordogna, G.; Carrara, P.; Pagani, M.; Pepe, M.; Rampini, A. Managing imperfect temporal metadata in the catalog services of Spatial Data Infrastructures compliant with INSPIRE. In Proceedings of the 2009 International Fuzzy Systems Association World Congress and 2009 European Society for Fuzzy Logic and Technology Conference, Lisbon, Portugal, 20–24 July 2009; pp. 915–920.
29. Lambe, P. *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*; Chandos Publishing: Oxford, UK, 2007; ISBN 9781843342274.
30. Burnett, K.; Ng, K.B.; Park, S. A comparison of the two traditions of metadata development. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 1209–1217. [CrossRef]
31. Working Group on Research Data—Science Europe (SE) Metadata. Available online: http://sedataglossary.shoutwiki.com/wiki/Metadata (accessed on 15 February 2022).
32. Data on the Web Best Practices Working Group—World Wide Web Consortium (W3C) Guidance on the Provision of Metadata. Available online: https://www.w3.org/2013/dwbp/wiki/Guidance_on_the_Provision_of_Metadata (accessed on 15 February 2022).
33. Harrison, M.; Farias Lóscio, B. Re: Intrinsic vs. Extrinsic Metadata (My Action #54). Available online: https://lists.w3.org/Archives/Public/public-dwbp-wg/2014Jul/0004.html (accessed on 15 February 2022).
34. Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
35. Research Data Alliance Metadata Standards Directory WG. Available online: https://www.rd-alliance.org/groups/metadata-standards-directory-working-group.html (accessed on 15 February 2022).
36. Research Data Alliance. Available online: https://www.rd-alliance.org (accessed on 15 February 2022).
37. Chen, S.; Alderete, K.A.; Ball, A. RDA Metadata Standards Directory. Available online: https://rd-alliance.github.io/metadata-directory/ (accessed on 15 February 2022).
38. World Wide Web Consortium (W3C) Semantic Web. Available online: https://www.w3.org/standards/semanticweb/ (accessed on 15 February 2022).
39. Zaveri, A.; Rula, A.; Maurino, A.; Pietrobon, R.; Lehmann, J.; Auer, S. Quality assessment for Linked Data: A Survey A systematic literature review and conceptual framework. *Semant. Web* **2016**, *7*, 63–93. [CrossRef]
40. American Society for Information Science and Technology Dublin Core™ Metadata Initiative. Available online: https://www.dublincore.org/collaborations/asist/ (accessed on 15 February 2022).
41. Weibel, S. The Dublin core: A simple content description model for electronic resources. *Bull. Am. Soc. Inf. Sci.* **1997**, *24*, 9–11. [CrossRef]
42. *ISO 15836-1:2017*; Information and Documentation—The Dublin Core Metadata Element Set—Part 1: Core Elements. International Standardization Organization (ISO): Geneva, Switzerland, 2017; p. 7.
43. *ANSI/NISO Z39.85*; The Dublin Core Metadata Element Set. National Information Standards Organization: Baltimore, MD, USA, 2012; p. 7.
44. *ISO 15836-2:2019*; Information and Documentation—The Dublin Core Metadata Element Set—Part 2: DCMI Properties and Classes. International standardization Organization (ISO): Geneva, Switzerland, 2019; p. 26.
45. Dublin Core™ Metadata Initiative DCMI Metadata Terms. Available online: https://www.dublincore.org/ (accessed on 15 February 2022).
46. Woodley, M.S. *Crosswalks, Metadata Harvesting, Federated Searching, Metasearching: Using Metadata to Connect Users and Information*; Getty Research Institute: Los Angeles, CA, USA, 2008; pp. 1–25.
47. Harpring, P. *Metadata Standards Crosswalks*; Getty Research Institute: Los Angeles, CA, USA, 2022.
48. Metadata Working Group of the Emory University Crosswalk of Core Metadata. Available online: https://metadata.emory.edu/guidelines/descriptive/crosswalk.html (accessed on 15 February 2022).
49. Leahey, A.; Barsky, E.; Brosz, J.; Garnett, A.; Gray, V.; Hafner, J.; Handren, K.; Harrigan, A.; Lacroix, C.; Pascoe, J.; et al. *Metadata for Discovery: Disciplinary Standards and Crosswalk Progress Report*; University of British Columbia Library: Vancouver, BC, USA, 2017.
50. Kaplan, S.; Garrick, B.J. On The Quantitative Definition of Risk. *Risk Anal.* **1981**, *1*, 11–27. [CrossRef]
51. *NORSOK Standard Z-013*; Risk and Emergency Preparedness Assessment. NORSOK: Lysaker, Norway, 2010.
52. *CAN/CSA-Q850-97 (R2009)*; Risk Management: Guideline for Decision Makers. Canadian Standards Association (CSA): Ottawa, ON, Canada, 1997.
53. De Marchi, B.; Ravetz, J.R. Risk management and governance: A post-normal science approach. *Futures* **1999**, *31*, 743–757. [CrossRef]
54. Aven, T.; Krohn, B.S. A new perspective on how to understand, assess and manage risk and the unforeseen. *Reliab. Eng. Syst. Saf.* **2014**, *121*, 1–10. [CrossRef]
55. Paltrinieri, N.; Dechy, N.; Salzano, E.; Wardman, M.; Cozzani, V. Towards a new approach for the identification of atypical accident scenarios. *J. Risk Res.* **2013**, *16*, 337–354. [CrossRef]
56. Paltrinieri, N.; Khan, F.; Amyotte, P.; Cozzani, V. Dynamic approach to risk management: Application to the Hoeganaes metal dust accidents. *Process Saf. Environ. Prot.* **2013**, *92*, 669–679. [CrossRef]
57. Kalantarnia, M.; Khan, F.; Hawboldt, K. Dynamic risk assessment using failure assessment and Bayesian theory. *J. Loss Prev. Process Ind.* **2009**, *22*, 600–606. [CrossRef]

58. Khakzad, N.; Khan, F.; Amyotte, P. Dynamic risk analysis using bow-tie approach. *Reliab. Eng. Syst. Saf.* **2012**, *104*, 36–44. [CrossRef]
59. Bucelli, M.; Paltrinieri, N.; Landucci, G. Integrated risk assessment for oil and gas installations in sensitive areas. *Ocean Eng.* **2018**, *150*, 377–390. [CrossRef]
60. Lee, S.; Landucci, G.; Reniers, G.; Paltrinieri, N. Validation of dynamic risk analysis supporting integrated operations across systems. *Sustainability* **2019**, *11*, 6745. [CrossRef]
61. Yang, X.; Haugen, S.; Paltrinieri, N. Clarifying the concept of operational risk assessment in the oil and gas industry. *Saf. Sci.* **2018**, *108*, 259–268. [CrossRef]
62. *JCGM 200:2012*; International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM). Joint Committee for Guides in Metrology (JCGM): Sèvres, France, 2012.
63. *ISO 10012:2003*; Measurement Management Systems—Requirements for Measurement Processes and Measuring Equipment. International Standardization Organization (ISO): Geneva, Switzerland, 2003; p. 26.
64. Dillon, R.L.; Paté-Cornell, M.E.; Guikema, S.D. Programmatic risk analysis for critical engineering systems under tight resource constraints. *Oper. Res.* **2003**, *51*, 354–370. [CrossRef]
65. Borgonovo, E.; Cillo, A. Deciding with Thresholds: Importance Measures and Value of Information. *Risk Anal.* **2017**, *37*, 1828–1848. [CrossRef] [PubMed]
66. Iwanaga, T.; Wang, H.H.; Hamilton, S.H.; Grimm, V.; Koralewski, T.E.; Salado, A.; Elsawah, S.; Razavi, S.; Yang, J.; Glynn, P.; et al. Socio-technical scales in socio-environmental modeling: Managing a system-of-systems modeling approach. *Environ. Model. Softw.* **2021**, *135*, 104885. [CrossRef] [PubMed]
67. Stefana, E.; Cocca, P.; Marciano, F.; Rossi, D.; Tomasoni, G. A review of energy and environmental management practices in cast iron foundries to increase sustainability. *Sustainability* **2019**, *11*, 7245. [CrossRef]
68. Paté-Cornell, M.E. Uncertainties in risk analysis: Six levels of treatment. *Reliab. Eng. Syst. Saf.* **1996**, *54*, 95–111. [CrossRef]
69. Raussand, M. *Risk Assessment—Theory, Methods, and Applications*; Wiley: Hoboken, NJ, USA, 2011; Volume 1, ISBN 978-0-470-63764-7.
70. Creedy, G.D. Quantitative risk assessment: How realistic are those frequency assumptions? *J. Loss Prev. Process Ind.* **2011**, *24*, 203–207. [CrossRef]
71. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv.* **2009**, *41*, 52. [CrossRef]
72. Wang, R.Y. A Product Perspective on Total Data Quality Management. *Commun. ACM* **1998**, *41*, 58–65. [CrossRef]
73. English, L.P. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*; Wiley: Hoboken, NJ, USA, 1999; ISBN 978-0-471-25383-9.
74. Eppler, M.J.; Helfert, M. A classification and analysis of data quality costs. In Proceedings of the International Conference on Information Quality, Cambridge, MA, USA, 5–7 November 2004; pp. 311–325.
75. Loshin, D. (Ed.) 4-Economic framework of data quality and the value proposition. In *Enterprise Knowledge Management*; Academic Press: Cambridge, MA, USA, 2001; pp. 73–99, ISBN 9780124558403.
76. Scannapieco, M.; Virgillito, A.; Marchetti, C.; Mecella, M.; Baldoni, R. The DaQuinCIS architecture: A platform for exchanging and improving data quality in cooperative information systems. *Inf. Syst.* **2004**, *29*, 551–582. [CrossRef]
77. Long, J.A.; Seko, C.E. A Cyclic-Hierarchical Method for Database Data-Quality Evaluation and Improvement. In *Information Quality*; Wang, R.Y., Pierce, E.M., Madnick, S., Fisher, C., Eds.; Taylor & Francis Group: New York, NY, USA, 2005; p. 15, ISBN 9781315703480.
78. Jeusfeld, M.A.; Quix, C.; Jarke, M. Design and Analysis of Quality Information for Data Warehouses. In Proceedings of the 17th International Conference on Conceptual Modeling, Singapore, 16–19 November 1998; pp. 349–362.
79. Aljumaili, M.; Karim, R.; Tretten, P. Metadata-based data quality assessment. *VINE J. Inf. Knowl. Manag. Syst.* **2016**, *46*, 232–250. [CrossRef]
80. European Space Imaging Our Satellites. Available online: https://www.euspaceimaging.com/ (accessed on 15 February 2022).
81. Van Etten, A. You Only Look Twice: Rapid Multi-Scale Object Detection In Satellite Imagery. *arXiv* **2018**, arXiv:1805.09512.
82. Kuželka, K.; Slavík, M.; Surový, P. Very High Density Point Clouds from UAV Laser Scanning for Automatic Tree Stem Detection and Direct Diameter Measurement. *Remote Sens.* **2020**, *12*, 1236. [CrossRef]
83. Dokic, T.; Kezunovic, M. Predictive Risk Management for Dynamic Tree Trimming Scheduling for Distribution Networks. *IEEE Trans. Smart Grid* **2018**, *10*, 4776–4785. [CrossRef]
84. Alabri, A.; Hunter, J. Enhancing the quality and trust of citizen science data. In Proceedings of the 6th IEEE International Conference on E-Science (E-Science 2010), Brisbane, QLD, Australia, 7–10 December 2010; pp. 81–88. [CrossRef]
85. Byabazaire, J.; O'Hare, G.; Delaney, D. Data quality and trust: A perception from shared data in IoT. In Proceedings of the IEEE International Conference on Communications, Dublin, Ireland, 7–11 June 2020. [CrossRef]
86. Khakzad, N.; Khakzad, S.; Khan, F. Probabilistic risk assessment of major accidents: Application to offshore blowouts in the Gulf of Mexico. *Nat. Hazards* **2014**, *74*, 1759–1771. [CrossRef]
87. Keßler, C.; de Groot, R.T.A. *Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap*; Lecture Notes in Geoinformation and Cartography; Elsevier: Amsterdam, The Netherlands, 2013; Volume 2013, pp. 21–37, ISBN 9783319006147.

88. Bovee, M.; Srivastava, R.P.; Mak, B. A conceptual framework and belief-function approach to assessing overall information quality. *Int. J. Intell. Syst.* **2003**, *18*, 51–74. [CrossRef]
89. Pacevicius, M.; Roverso, D.; Rossi, P.S.; Paltrinieri, N. Risk of crack formation in power grid wooden poles and relationship with meteorological conditions: A Norwegian case study. In Proceedings of the Safety and Reliability—Safe Societies in a Changing World, Trondheim, Norway, 17–21 June 2018; pp. 1527–1533, ISBN 978-0-8153-8682-7.
90. Alhelou, H.H.; Hamedani-Golshan, M.E.; Njenda, T.C.; Siano, P. A survey on power system blackout and cascading events: Research motivations and challenges. *Energies* **2019**, *12*, 682. [CrossRef]
91. Pacevicius, M.; Dammann, D.O.; Gazzea, M.; Sapronova, A. Heterogeneous Data-merging Platform for Improved Risk Management in Power Grids. In Proceedings of the 67th Annual Reliability and Maintainability Symposium, Orlando, FL, USA, 24–27 May 2021; p. 7. [CrossRef]
92. Bucelli, M.; Landucci, G.; Haugen, S.; Paltrinieri, N.; Cozzani, V. Assessment of safety barriers for the prevention of cascading events in oil and gas offshore installations operating in harsh environment. *Ocean Eng.* **2018**, *158*, 171–185. [CrossRef]
93. Marler, R.T.; Arora, J.S. The weighted sum method for multi-objective optimization: New insights. *Struct. Multidiscip. Optim.* **2010**, *41*, 853–862. [CrossRef]
94. Triantaphyllou, E. Multi-Criteria Decision Making Methods. In *Multi-Criteria Decision Making Methods: A Comparative Study*; Applied Optimization; Springer: Boston, MA, USA, 2000; Volume 44, p. 16. [CrossRef]

Article VIII - Data-informed Risk Analysis of Power Grids: Application of Method for Managing Heterogeneous Datasets

Pacevicius, M., Ramos, M., Roverso, D., Thun Eriksen, C., Paltrinieri, N., 2022b. **Data-informed Risk Analysis of Power Grids: Application of Method for Managing Heterogeneous Datasets**. ASME J. Risk Uncertainty Part B. p. 44.

This article is awaiting publication and is not included in NTNU Open

Article IX - Heterogeneous Data-merging Platform for Improved Risk Management in Power Grids

Pacevicius, M., Dammann, D.O., Gazzea, M., Sapronova, A., 2021a. **Heterogeneous Data-merging Platform for Improved Risk Management in Power Grids**, in: The 67th Annual Reliability and Maintainability Symposium. Orlando, p. 7.

Heterogeneous Data-merging Platform for Improved Risk Management in Power Grids

Michael Pacevicius, MSc, Norwegian University of Science and Technology (NTNU) / eSmart Systems

Dyre Oliver Dammann, PhD, University of Alaska Fairbanks

Michele Gazzea, MSc, Western Norway University of Applied Sciences (HVL)

Alla Sapronova, PhD, StormGeo AS

Key Words: Heterogenous data merging, Platform development, Dynamic risk analysis, Power grids, Remote Sensing.

SUMMARY & CONCLUSIONS

Vegetation-related and wind-induced outages are two primary causes for disruption in electricity transmission and distribution networks. Wind can either be a direct cause of disruption, e.g. by over-constraining infrastructures, leading to cracks and ruptures, or indirectly, e.g. when leading to trees falling on power lines. The ability to efficiently address such hazards remains limited, despite the considerable increase in relevant data over recent years to help mitigate risk. Moreover, to our knowledge, no platform exists that addresses wind and vegetation hazards in combination for power-grid management. The development of such a solution is a timely exercise, as recent and predicted climate change can lead to more extreme weather events and difficulties to forecasting them.

In this paper, we develop and present a platform to elevate situation awareness in the context of vegetation- and wind-induced risk near power lines, which can improve power network resilience and reduce cost of electricity outages. In this process, we demonstrate how heterogeneous data sources with variable temporal and spatial scales (e.g. satellite imagery, ground surveys, outage reports, and weather data) can be exploited to generate end-user-oriented applications.

We apply the developed strategy to a case study in western Norway, where we develop a support product to assist grid operators. The final product can enhance reliable vegetation status awareness and predictions of weather-related outages to enable effective resource allocation and ensure safety of ground crews.

1 INTRODUCTION

Transmission System Operators (TSOs) and Distribution System Operators (DSOs) face a multitude of challenges when attempting to maximize the uptime of power grids and meet increasing power demand on an aging infrastructure. Among those challenges are outages induced by vegetation and wind [1], which are primary causes for outages throughout the word [2,3]. Furthermore, climate change is expected to increase both the intensity and the frequency of extreme weather events [4], which will likely increase the number of outages.

Adequately addressing such challenges is difficult due to the large spatial scales of power grids, regional differences, and temporal timespan of observations. Any approach to address those challenges thus rests on effective use of data that can resolve the relevant spatiotemporal scale. Wind data is for example relevant on a timescale ranging from hours to days and km-scale spatial resolution. Vegetation on the other hand, changes less rapidly than weather and is relevant on a timescale ranging from months to years. However, the spatial variability in vegetation is on the meter- to sub-meter scale as opposed to wind. The relevance of scale for parameters impacting decision-making by grid operators is illustrated in Figure 1.

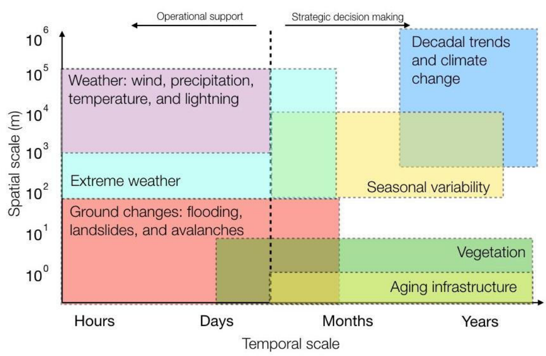


Figure 1 – Relevance of spatial and temporal scale for information support to power system operators

Responsive solutions enabling TSOs and DSOs to obtain a real-time awareness over large-scale infrastructures distributed across regions and countries are critical. Yet, current inspection approaches are often infective, reducing that way the number of inspections possible. Additionally, the management of the acquired data is usually suboptimal, partly due to the difficulty in gathering and technically combining relevant information from heterogenous data sources within a suitable environment [5]. As a consequence, the analysis of such data can result in a partial or inadequate risk picture.

Replacing conventional risk assessment methods with more automatized and data-driven approaches is needed to enable efficient support for maintenance operations. Remote sensing represents a powerful tool for that purpose, enabling large-scale cost-effective surveys. This has opened up for a range of new vegetation monitoring solutions [6–8]. The use of state-of-the-art weather modelling has furthermore also been used for reliability assessment of power grids [9–11]. Here, we present an effort to combine remote sensing-based vegetation assessment with wind-based risk analyses. To our knowledge, this has not previously been attempted. However, we argue that this is a necessary approach based on the strong link between wind and vegetation in outage events. Additionally, we also argue that automatization strategies for adaptive data integration need to be defined in order to enable dynamic risk analyses [12].

We first describe the strategy chosen to develop the architecture of the platform and the data that was considered for addressing (1) the presence of vegetation along power lines and (2) the impact of wind-related outage prediction within the context of power-grid management (Section 2). The application of the strategy is then demonstrated in Section 3. Finally, we discuss the developed approach in the context of current solutions, user needs, and future development (Section 4).

2 IDENTIFYING USER AND DATA NEEDS

2.1 *Identifying user needs*

This work was carried out in close collaboration with Sogn og Fjordane Energi (SFE) - a Norwegian power grid company, which supported the research team by providing data, recommendation, and feedback through all parts of the project. A critical component of the collaboration with SFE was the ability to determine key user needs within the main objective - to maximize operational uptime for large-scale power grids in potentially critical environmental conditions.

We started, based on [13], by developing a framework to help guide the process of translating user-specific needs into a monitoring system that incorporates (1) identification of user needs, (2) data acquisition, and (3) method development (Cf. Table 1).

First, we identified the principal user-needs (step 1a in Table 1) as:

(1) cost effectively identify locations of trees encroaching on power line corridors to aid vegetation management planning and identification of grid vulnerabilities; and
(2) predict local outages based on forecasted weather conditions to assess short-term vulnerability and help distribute manpower and/or redirect power in the network.

Second, we identified the most relevant parameters for the identified needs (step 1b in Table 1) to be (1) vegetation above 4.5 meters that exist within 6 meters from the power line and (2) wind speed and direction. These conclusions were based on both thresholds defined by regulations [14] and close exchanges with experienced personnel in the industry.

Third, we explored the spatiotemporal scale of the major relevant parameters for operational and strategic support to power system operators (Figure 1) (step 1c in Table 1).

Forth, we explored and evaluated relevant data in light of user needs and spatial and temporal scales, including internal SFE data specific to the power grid's infrastructures (e.g. component data, historical outage data, ground survey data), external local information (e.g. topography, grid data), and globally available data (e.g. satellite and model data outlined in Section 2.2) (steps 2a, 2b, 2c of Table 1).

Table 1 – Step-by-step Actions for Platform Development

I. Defining User Needs
I.a) Determine specific user needs in terms of monitoring and prediction for operational and strategic decision support. **I.b)** Determine intrinsic variables and environmental parameters relevant for the specified need, user, and area. **I.c)** Determine appropriate spatiotemporal scales and required accuracy for the chosen approach based on user needs.
II. Data Evaluation & Acquisition
II.a) Acquire internal user data based on needs and availability (e.g., power grid data, field surveys). **II.b)** Acquire external local data based on needs (e.g., maps of vegetation species and power-infrastructure). **II.c)** Acquire global data (e.g., atmospheric model and satellite data) & models/methods relevant for the desired approach and scale.
III. Method Development & Integration
III.a) Develop and adapt models and methods for tracking and support of task-specific needs. **III.b)** Combine task-specific results for addressing zeroth order user needs. **III.c)** Assess and validate results and make adjustments according to end-user feedback.

2.2 *Data*

To assess vegetation height, we utilized airborne LiDAR data acquired during power line inspections and provided by SFE (Figure 2a). Vegetation height information was extracted from relevant classes in the point clouds (3 dimensions) and projected on a 2-dimensional raster (Figure 2b, 2c, 2d). For large km-scale vegetation assessment we utilized optical and infrared satellite data from WorldView-2 and Pleiades-1. These two systems were chosen as they enable global coverage with 0.5-meter resolution - high enough to resolve individual trees. The 2-dimensional raster extracted from the LiDAR data was then superimposed on the satellite images, as is illustrated in Figure 2e.

Wind data was acquired from the European Center for Medium-range Weather Forecast (ECMWF) ERA5 reanalysis [15]. This data includes hourly wind speed and direction on a 30 km grid.

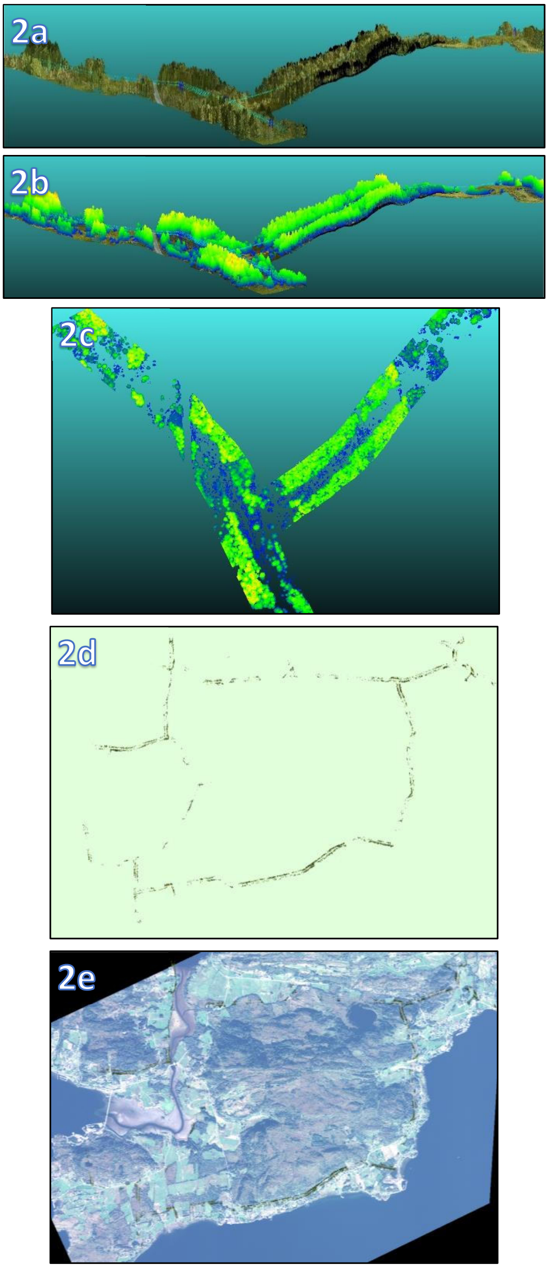


Figure 2 (a to e) – Combination of LiDAR data and satellite images. High vegetation is isolated in the point cloud and projected on satellite images to be used as ground-truth for machine learning tasks.

3 THE GRIDEYES PLATFORM

In response to the third step in the proposed framework, we developed an *Analytical pipeline* (step 3a in Table 1) describing the analytical steps of the GridEyeS platform (Table 2 and Figure 3). This pipeline illustrates two initially separate paths to develop monitoring capabilities in respect to vegetation (Path 1) and wind (Path 2) in response to the two primarily identified user needs. Steps 1-4 in both Path 1 and 2 (Table 2) were carried out individually, resulting in information that can directly aid in prediction of wind and vegetation related outages respectively. We combined the two assessments in a final fusion (step 3b in Table 1; step 6 in Table 2 and Figure 3).

As an answer to the need for automatization strategies enabling adaptive data integration, we furthermore suggest that the outlined pipeline can be integrated into a general system hosting the analytics and enabling an automatic processing of the information (i.e., the GridEyeS platform). The system can be divided into three main blocks: (1) data acquisition, (2) data processing, (3) distribution of results (Figure 4). The ownership of the data used in the platform can be split into two categories: external (i.e., acquired via third parties) and internal (i.e., provided by stakeholders involved in the platform development). Once the process for data acquisition has been clarified and the relevant *Data Sources* belonging to these two categories are connected to the developed solution, their content is transmitted to the *Processing Unit*. The *Processing Unit* is itself a combination of three modules: the *Analysis & Decision* module (consisting in the *Analytical pipeline* previously detailed), the *Retrain, Adapt & Update* module (principally covering the evolutionary process of updating the *Analytical pipeline* based on needs and further data acquisitions) and the *Archiving* module (ensuring that results are stored and evolutions of the solutions or detected patterns tracked over time for future analysis work or data exploitation). The output of the *Processing Unit* can thereafter be forwarded to the *Communication* block, ensuring that the right information is forwarded in time, to the right person or device and in the right format.

Further links between the main blocks can also be identified, such as a direct connection between the *Data Sources* and the *Dashboard* present in the *Communication* block. The link between the *Processing Unit* and the *Data Sources* additionally shows the possibility for modules of the *Processing Unit* to directly feed existing databases. A feedback loop originating from the *Dashboard* finally gives the possibility for the user to send information requests to both the *Data Sources* and the *Processing Unit* directly and on-demand. An illustration of the global platform architecture is given in Figure 4.

Table 2 – Elements of the Analytical Pipeline

<table>
<tr><td colspan="2"><u>First Step</u>: (0) Survey request for specific region.</td></tr>
<tr><td><u>Path 1 - Threatening Tree Detection</u>

(1.1a) Acquire optical satellite imagery.
(1.1b) Acquire aerial imagery and classified LiDAR point-clouds.
(2.1a) Pre-process satellite-image (vegetation index generation, texture classification, etc.)
(2.1b) Subsection data by defining AoI (Areas of Interest – here, power-line corridors).
(3.1) Generate satellite-based tree classification. (Achieved by using 2D projection of the “high vegetation” class in the LiDAR point-cloud and using it as ground truth for vegetation detection on the satellite images. [16,17])
(4.1.a) Pinpoint sections of the grid with trees too close to the power lines.
[(4.1.b) Estimate total number of trees].

[(5.1) Tree-related likelihood-of-outage estimation based on distance to trees, density, types and surrounding terrain].</td><td><u>Path 2 - Weather-related Threat level Forecasting (Initial Focus on wind speed & direction)</u>

(1.2.a) Acquire historical wind data (ERA5 reanalysis).
(1.2.b) Acquire wind predictions from ECMWF global forecast model.
(1.2.c) Acquire historical outage data reported for the power grid.
(2.2a) Extract relevant variables from model data for AoI.

(2.2b) Link historical outages to specific locations and transformers.
(3.2a) Establish distribution of wind speed and direction throughout study area based on historical data.
(3.2b) Determine specific wind conditions during historical outages.
(3.2c) Establish a likelihood-of-outage function based on the correlation between outages and wind conditions.
(4.2.a) Forecast wind-based likelihood-of-outage based on locally derived likelihood-of-outage function and current weather forecast.
[(4.2.b) Determine likelihood-of-outage variability in different regions based on historical weather data].
[(5.2) Estimate weather-related likelihood-of-outage based on additional weather parameters including precipitation, temperature, and lightning].</td></tr>
<tr><td colspan="2"><u>Final Step</u>: (6) Merge vegetation analysis and weather-related information for final combined likelihood-of-outage estimation.</td></tr>
<tr><td colspan="2"><u>Outputs</u>:
Output 1: Output for costs estimations of general clear-cutting operations.
Output 2: Output for resource managers and grid operators in control rooms in prevision of storm.
Output 3: Output for likelihood-of-outage mapping per region supporting resource prioritization, upgrading of infrastructure and preparation for extreme scenarios.</td></tr>
</table>

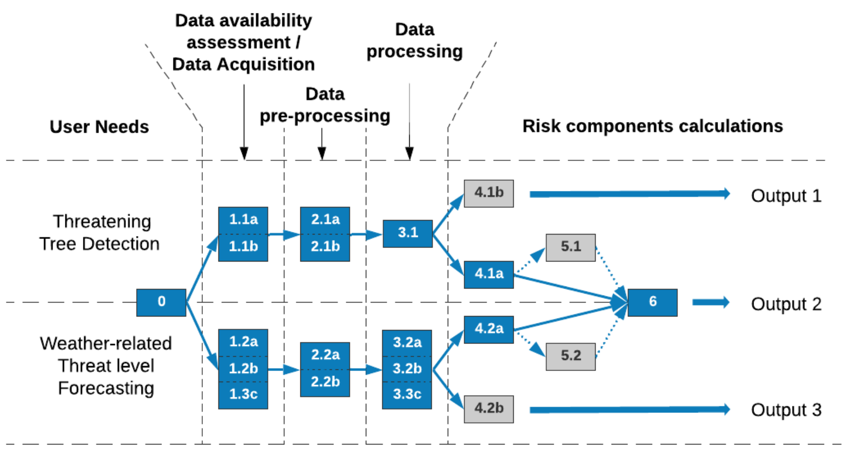


Figure 3 – Elements Assembling in the Analytical Pipeline.
Note: Doted arrows represent links to steps soon to be used for reinforcements of results provided in step 6.

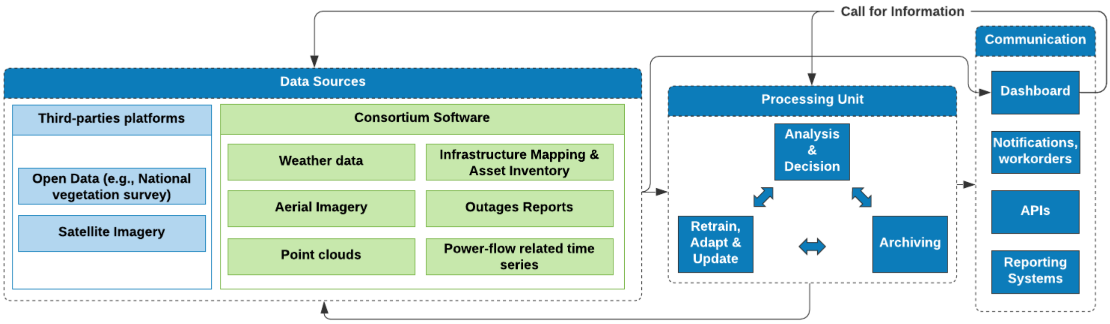


Figure 4 – High-level computational architecture of the GridEyeS product

4 DISCUSSION

4.1 *Results and applications*

Through this work, we demonstrate how an information platform can be tailored to deliver key services to end-users while considering both technological advances and important constraints related to heterogeneous data sources. The GridEyeS platform suggested here in the context of power-grid management can be used to obtain a better awareness of the grid status with regards to vegetation-related risk near power lines. The platform also provides insights into wind-related risk by assessing the wind conditions during past outage events and projecting this onto a weather forecast.

Preliminary promising results were presented to SFE and include heatmaps highlighting areas which are likely to face disturbances due to vegetation encroachment in proximity of power lines. Figure 5 reports an example of such results for the area shown in figure 2e, using a *Tree Density Index* (ranging from 0 to 1) that is based on both the proximity and the density of vegetation in the surrounding of the power lines.

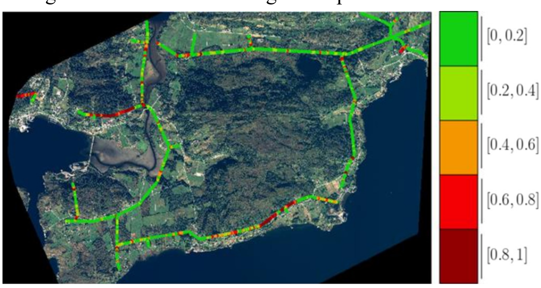


Figure 5 – Heatmap using a Tree Density Index to indicate high and low exposition of power lines to trees.

In addition, results also include areas prone to wind-related disturbances, using a likelihood-of-outage function directly derived from wind predictions, as shown in figure 6. More details relative to the quantitative achievements of the platform can be found in [16] and [17].

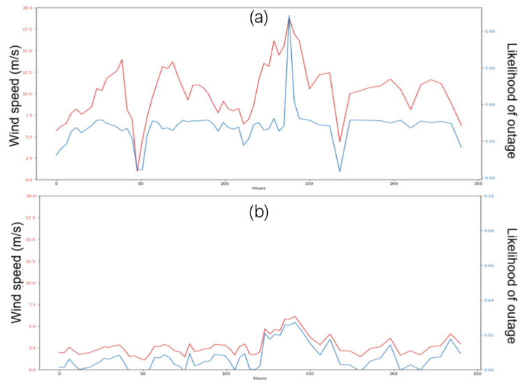


Figure 6 – 10-days wind forecast January 5-15 (red) during the storm "Nina" hitting the west coast of Norway in 2015 for two locations: Askrova (a), and Osstrupen-Steindal (b). The blue curves represent a likelihood-of-outage function derived from correlations between outages and wind-speed in the area under review in the period 1999-2019.

The results from this work may be used to identify sections of high-risk areas and limit the number of weak points in the grid (e.g., section of exposed unstable trees or old infrastructure) and/or to take adequate preventive actions (e.g., schedule larger operation crew in exposed areas based on a storm forecast or redirecting power in the network). By natively including a *Retrain, Adapt & Update* module in the *Processing Unit* of the solution, we furthermore support the use of dynamic risk analysis approaches, able to cope with situation modifications and appropriately restructuring the risk calculation process when facing circumstance changes [18]. This enables a realistic risk depiction, optimal decision making, and better responsive actions to minimise number of outages in power grids and their impacts.

4.2 *Improvement and future development*

The GridEyeS platform is under development as a collaboration project between two companies - eSmart Systems and StormGeo, funded by the European Space Agency (ESA). The planned completed platform will include (1) an additional 5th step which considers additional parameters and analyses for the reinforcement of results provided in step 6 and (2) expand the 4th step with steps 4.1b and 4.2b, incorporating more detailed analyses to support further resource management tasks for which grid operators are responsible (Table 2 and Figure 3).

Other possible improvements include incorporating additional data and processing methods to increase the pertinence of the provided predictions. Such sources can be - but are not limited to-: asset information (transformer age, substation type, etc.), synthetic aperture radar imagery, forest classification maps, topographic data, power flow time series, and demographic data. This will especially enable addressing dimensions not considered in the platform development so far, but addressed in other monitoring solutions, such as tree height, species categorization and growth estimation, forest fire risks, etc., which are as many relevant parameters also known to influence the risk levels in power grids.

5 CONCLUSIVE REMARKS

Traditional risk management approaches have the potential to strongly benefit from an increased access to new types of data sources. This is especially favoured by the global trends enabling data acquisition with increased quality, higher spatiotemporal resolution and at lower costs. At the same time, this leads to further challenges in terms of heterogeneous data merging and requires both cross-disciplinary expert collaboration and the existence of an environment enabling to host the analytics to be made. The present paper addresses this problematic and describes the preparation of a platform supporting these operations. More especially, we showed how we could merge data from heterogeneous data sources in the context of power-grid management and deliver insights relative to multiple hazards simultaneously into one tool. Our solution indicates in this manner how new types of data sources (especially remote sensing-based data) can be used by operators in a more efficient and dynamic way, enabling a better risk management of power grids. Finally, the described application is also a good illustration of the usefulness of the suggested frameworks in operational projects in general.

ACKNOWLEDGMENTS

This work is funded by the European Space Agency (ESA) (GridEyeS, project No. 4000127831/19/NL/MM/ra), which the authors would like to thank for their support. The authors also thank Sven Arild Kjerpeset and Kristian Vassbotten at SFE; Torleif Markussen Lunde at StormGeo; Reza Arghandeh at the Western Norway University of Applied Sciences; and Davide Roverso at eSmart Systems for valuable advice and technical support. WorldView-2 and Pleiades-1 satellite imagery were acquired from Apollo Mapping.

REFERENCES

1. Eggum E. *Rapport Nr. 29-2019 - Avbrotsstatistikk 2018*. Oslo, 2019 : 101 p.
2. Sforna M, Delfanti M. Overview of the events and causes of the 2003 Italian blackout. *2006 IEEE PES Power Systems Conference and Exposition, PSCE 2006 - Proceedings*. 2006 : 301–308.
3. Haes Alhelou H, Hamedani-Golshan M, Njenda T, *et al.* A Survey on Power System Blackout and Cascading Events: Research Motivations and Challenges. *Energies* 2019 ; 12 : 682.
4. Masson-Delmotte V, Pörtner H-O, Skea J, *et al. Global Warming of 1.5°C: Summary for Policymakers*. Geneva, Switzerland, 2018 : 32 p.
5. Pacevicius M, Roverso D, Salvo Rossi P, *et al.* Smart Grids : Challenges of Processing Heterogeneous Data for Risk Assessment. *Probabilistic Safety Assesment and Management - PSAM 14*. Los Angeles, 2018 : 11.
6. MAXAR - COTESA. Automated Powerline Monitoring. 2020.
7. LiveEO. Infrastructure Monitoring From Space. 2020.
8. Terramonitor. Reshape your view on the world. 2020.
9. Dunn S, Wilkinson S, Alderson D, *et al.* Fragility Curves for Assessing the Resilience of Electricity Networks Constructed from an Extensive Fault Database. *Nat. Hazards Rev.* 2018 ; 19 : 1–10.
10. Solheim OR, Kjolle G. Wind dependent failure rates for overhead transmission lines using reanalysis data and a Bayesian updating scheme. *2016 Int. Conf. Probabilistic Methods Appl. to Power Syst. PMAPS 2016 - Proc.* 2016 ;.
11. Panteli M, Pickering C, Wilkinson S, *et al.* Power System Resilience to Extreme Weather: Fragility Modeling, Probabilistic Impact Assessment, and Adaptation Measures. *IEEE Trans. Power Syst.* 2017 ; 32 : 3747–3757.
12. Villa V, Paltrinieri N, Khan F, *et al.* Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. *Saf. Sci.* 2016 ; 89 : 77–93.
13. Dammann DO, Eicken H, Mahoney AR, *et al.* Assessing sea ice trafficability in a changing arctic. *Arctic* 2018 ; 71 : 59–75.
14. Haaverstad O, Gåserud Ø. *Skogrydding i kraftledningstraséer*. Oslo, 2016 : 27 p.
15. European Centre for Medium-Range Weather Forecasts. ERA5. 2020 ;.
16. Pacevicius M, Gazzea M, Dammann DO, *et al. GridEyeS Deliverable D3: Technical Feasibility Assessment*. 2020 : 91 p.
17. Gazzea M, Pacevicius M, Dammann DO, *et al.* Vegetation Monitoring along Power Lines using Satellite Imagery and Machine Learning. 2020 ; 1–10.
18. Paltrinieri N, Khan F. *Dynamic Risk Analysis in the Chemical and Petroleum Industry: Evolution and Interaction with Parallel Disciplines in the Perspective of Industrial Application*. Elsevier, 2016 : 239 p.

BIOGRAPHIES

Michael Pacevicius
Department of Mechanical and Industrial Engineering
Norwegian University of Science and Technology NTNU
Richard Birkelands vei 2B
7034 Trondheim, Norway

e-mail: michael.f.pacevicius@ntnu.no

Michael Pacevicius is an industrial PhD candidate in the RAMS group at NTNU, Norway and works as a researcher in eSmart Systems, a company delivering high-tech IT solutions for power grid related companies. His research activities focus on the development and implementation of dynamic risk analysis methods for large-scale interconnected power systems. He has a MSc. in Operational safety, Risks and Environment from the Université de Technologie de Troyes (UTT) in France and a MSc. in Economics and Business Administration from the Technische Universität Braunschweig (TUBS) in Germany. He worked as a project coordinator and analyst in the Big Data business development department of SAP in Munich, Germany, before joining eSmart Systems back in 2017.

Dyre Oliver Dammann, PhD
Department of Geosciences
University of Alaska Fairbanks
2156 Koyukuk Dr.
Fairbanks, Alaska, USA

e-mail: dodammann@alaska.edu

Dyre Oliver Dammann is a researcher at the University of Alaska Fairbanks focusing on remote sensing application development relevant to trafficability and infrastructure primarily in the context of sea ice. His background includes a Ph.D. focusing on developing methods using synthetic aperture radar and a M.S. in large-scale atmospheric modelling.

Michele Gazzea
Department of Computer Science and Electrical Engineering
Western Norway University of Applied Sciences
Inndalsveien 28
5063 Bergen, Norway

e-mail: michele.gazzea@hvl.no

Michele Gazzea is a Ph.D. candidate at Western Norway University of Applied Sciences in Bergen (Norway). His current research interests are in data analytics, machine learning, computer vision, and remote sensing applications. He received his Bachelor degree in Information Engineering and his Master degree in Automation and Control Engineering, both from the University of Padova (Italy). The studies were mainly focused on control theory and applications, robotics, and computer vision. He worked for one year in Cielle S.r.l. in Treviso (Italy) as an R&D engineer studying and designing diagnostic techniques on milling and engraving machines. He worked as a researcher for Electrolux to perform model-based analysis of power consumption in washing machines until he started his new occupation as a Ph.D. student.

Alla Sapronova
StormGeo AS
Nordre Nøstekaien 1
5011 Bergen, Norway

e-mail: alla.sapronova@stormgeo.com

Alla Sapronova is Data Scientist with 10+ years of experience in applied data science, process modelling, R&D activities and software development. She has a PhD in Physics and Mathematics. She delivered AI/ML solutions (including deep learning, big data analysis, etc.) within digitalization, machine learning applications and data driven decision support for renewable energy sector, electricity consumption forecast, fishing industry, shipping and logistics. In 2018-2019 she received support from Norwegian Research Council as inventor and primary investigator in two innovation projects within digitization and data science application for fish farming and renewable energy. She is currently Lead Data Scientist at StormGeo AS.

Article X - Addressing the Importance of Data Veracity during Data Acquisition for Risk Assessment Processes

Pacevicius, M., Paltrinieri, N., Thieme, C.A., Salvo Rossi, P., 2021b. **Addressing the Importance of Data Veracity during Data Acquisition for Risk Assessment Processes**, in: The 67th Annual Reliability and Maintainability Symposium. Orlando, p. 7.

Addressing the Importance of Data Veracity during Data Acquisition for Risk Assessment Processes

Michael Pacevicius, MSc, Norwegian University of Science and Technology (NTNU) / eSmart Systems
Nicola Paltrinieri, PhD, Norwegian University of Science and Technology (NTNU)
Christoph Alexander Thieme, PhD, Norwegian University of Science and Technology (NTNU)
Pierluigi Salvo Rossi, PhD, Norwegian University of Science and Technology (NTNU)

Key Words: Data veracity, Dynamic risk management, Confirmation factor, Power grids.

SUMMARY & CONCLUSIONS

The veracity of information (i.e., its quality of being and remaining true, accurate, and complete) is a pillar of efficient risk management. The informative capacity of the data on which the risk management process relies needs to be fully kept across the entire information pipeline in order to ensure that risk can be properly understood and managed. Unfortunately, research shows that the informative capacity of data may - partially or entirely - be lost between the generation and the final use of a piece of information. This problem starts with the capture of information, where inconsistencies may already be observed between the reality of a phenomenon and the data supposedly reporting its measurement. As a consequence, this can lead to inadequate decision making when answering a risky event and, thus to a critical escalation of the situation. Such circumstances have been reported as contributing factors in several well-known large-impact accidents (e.g., *Three Mile Island*, 1979; *BP Texas City Refinery*, 2005; *Deepwater Horizon*, 2010) and continue to be faced in high-risk infrastructures nowadays.

The multiplication of information sources made available through advances in the Internet of Things (IoT) and digital fields offers an opportunity to address this issue, as more and more data sources can be used to confirm a single fact. That way, decision-makers can better detect inconsistencies in the data used for risk analyses and apply appropriate corrective actions. However, this comes with several challenges. Firstly, conventional risk management approaches need to be rethought and restructured to enabling a dynamic updating of the risk picture as new information is made available. Secondly, they need to enable a characterization of the information quality by providing details on the level of uncertainties related to the generated risk picture. Thirdly, the data capture process needs to be properly understood in order to ensure that possible data corruption modes are correctly identified.

This paper discusses the points above by focusing on the veracity of information during the capture of data for risk assessment purposes. We discuss how multiple data sources may be managed to reduce uncertainties in this phase. A case study on the presence of vegetation close to power lines illustrates the related implications.

1 INTRODUCTION

Common risk assessments remain mostly focused on the processes themselves, assuming the existence of a reliable supporting infrastructure [1]. However, the information pipeline responsible for transmitting a piece of data from phenomenon observation to decision making (data capture, data transmission, data pre-processing, information processing, results transmission) represents a complex system of systems, which all can be a source of data corruption eventually leading to an inadequate decision making. Corruption is here understood as the possibility for a piece of information to lose its veracity, i.e., its quality of being and remaining true, accurate, and complete. Different major accidents can be used to illustrate the important consequences of a degraded information management process. In the *Three Mile Island* accident (1979), decision-makers have built their reasoning and took action in an emergency situation based on inaccurate information, being informed that automatic safety procedure had been successfully executed, while they were not [2]. Misreading of pressure information and ignoring of warnings about cement weaknesses were also some of the root causes responsible for the *Deepwater Horizon* catastrophe (2010), the biggest offshore oil spill in US history [3]. Additional events happening in between the *Three Mile Island* accident and the *Deepwater Horizon* catastrophe (e.g., the *BP Texas City Refinery* accident in 2005 [4]) are other illustrations of accidents showing that a piece of information may:

- not have been generated (e.g., sensors not working), and/or
- have been generated in an inappropriate way (e.g., sensors generating false information), and/or
- have been wrongly transported and distributed (i.e., telecommunication network failure), and/or
- have been treated by inappropriate analysis methods (e.g., outdated algorithm utilization), and/or
- have been wrongly interpreted by operators (e.g., human errors: making wrong decisions despite receiving the correct information in the right format).

There is thus a need to increase attention on the gathering, transmission, and processing of information to ensure higher reliability of risk management processes. This topic becomes even more relevant in today's era of big data, where more and more digital information is made available and considered for risk management. Although the veracity of data is discussed in the literature for different applications, implications from the perspective of risk assessment have not yet been thoroughly examined. The present paper addresses thus this topic during the first step of the information chain used for risk assessment - data acquisition - through:

(1) the understanding of the implications and adaptation requirements for conventional risk approaches to fully integrate the veracity dimension in the risk assessment process; and
(2) discussing a framework for enabling the management of existing data sources so that a more reliable risk assessment can be executed.

This paper first revisits risk fundamentals in Section 2, clarifying the concept of Dynamic Risk Management (DRM), and highlighting the importance of reliable data acquisition for this purpose. It then builds on approaches typically used for data validation to integrate the veracity of information into conventional risk assessment approaches in Section 3. Finally, Section 4 presents a case study on vegetation management for power grids risk assessment. The presence of vegetation in the surrounding of power lines significantly impacts the probability of outage in power grids, and power grid outages can have on serious impact on modern societies.

2 FUNDAMENTALS

2.1 Risk Definition

One of the most renowned definitions of risk was given by Kaplan and Garrick [5]. It states that risk (R) can be expressed by what can go wrong (scenario s), what likelihood it will have (probability p), and how severe consequences will be (consequence c):

$$R = f(s, p, c) \tag{1}$$

On the other hand, several review articles [6–8] collect parallel risk definitions from the scientific literature to demonstrate the multiplicity of perspectives on the understanding of the concept of risk. In [9], the risk is defined as an uncertain consequence of an event or an activity with respect to something that humans value. For [10], Risk equals the expected loss. And for [11], the risk is the potential for realization of unwanted, negative consequences of an event.

There is thus not one single approach, but several paths leading to relatively different results, which may be all beneficial but intrinsically incomplete. This is demonstrated by the occurrence of major accidents whose scenarios were disregarded by safety reports because being deemed improbable [12]. As an attempt to provide a more comprehensive risk definition, Aven and Krohn (2014) suggest including knowledge (k) as a new dimension in the original definition (1):

$$R = f(s, p, c, k) \tag{2}$$

Here, this definition is retained due to the strong overlap existing between the concepts of *knowledge* and *veracity*. Its use is further detailed in Section (3).

2.2 Risk Management

Several examples of frameworks addressing risk management or governance may be found in standards and related contributions in literature [14]: i) "Risk management: guideline for decision makers" by the Canadian Standard Association (standard CSA Q850-97) [15]; ii) "Risk management: principles and guidelines" by the International Organization for Standardization (standard ISO 31000:2018) [16]; iii) "Risk governance framework" by the International Risk Governance Council [9]; and "Risk and emergency preparedness assessment" by the Norwegian petroleum industry (standard NORSOK Z-013) [17].

The mentioned risk management frameworks unanimously address the following steps: pre-assessment, risk assessment, tolerability/ acceptability judgment, risk management, and risk communication. Treatment of uncertainties is also emphasized, and different related practices are suggested. ISO 31000 defines risk as uncertainty to achieve an objective [16]. The IRGC framework [9] distinguishes between uncertainty and ambiguity. Uncertainty refers to a lack of clarity over the scientific or technical basis for decision making, whereas ambiguity gives rise to several meaningful and legitimate interpretations of accepted risk assessment results. Ambiguity may refer to potential different values leading to a variety of interpretations.

Uncertainties can arise at different levels and moments of the risk management process, and they may be related to data, models, or the decision-making phase. Most of the risk management frameworks invite to consider and acknowledge all forms of uncertainties, not only technical but also social. Decision making under uncertainty usually relies on the consideration and comparison of multiple scenarios, therefore requiring continuous improvement to maximize the likelihood of appropriate judgments. For instance, in the presence of emerging risks, differences among actual and expected results are likely, due to limits in experience and knowledge. The introduction of continuous improvement is thus fundamental in order to proceed towards effective and efficient risk management.

Constant monitoring supports continuous improvement, which is already a recurrent step in the risk management frameworks presented. However, there are very few references on what and how to monitor or measure. The monitoring process is often related to the level of achievement of objectives or to the adequacy of assumptions with observed consequences. While on the contrary, the IRGC framework insists on the monitoring of [9]: equity in the repartition of risks and benefits among different categories of populations; and transparency and availability of information for various stakeholders.

2.3 Dynamic Risk Management

DRM has an evident focus on the concept of risk. However, one should not be misled into thinking that this domain aims at managing "dynamic risk." Analogously to dynamic risk analysis [18], it refers to the management process that is designed to dynamically handle the risk of a system. One important role of DRM is the conversion of the traditionally static process of risk analysis into a dynamic technique with the capacity to be regularly updated. This should include a common understanding of tolerability and acceptability of risk levels and a clear operational scope [19]. Despite its relevance, the frequency of updates of the process falls outside the purely methodological scope of this research [18]. DRM refers to the risk management frameworks (DRMF) designed to be dynamic, which optimally enables restructuration, updates, and iterations when needed. Whether a DRMF updates instantaneously, yearly, or every decade, it will always keep its dynamic characteristic as it is independent of the actual use. On the contrary, non-DRM still has the possibility to be updated but with considerable inertia leading it to require substantial efforts, time, and energy by appointed teams of experts and managers.

An illustrative example can be made by means of computer software coding. When a software code is meant to be updated based on the inputs provided while running, its structure and characteristics should be defined and designed accordingly. One can say that this software code is dynamic. When, instead, a software code does not accept any input for the sake of updates while running, the only option is modifying the lines of code that we want to be changed. We can define the latter code as static. Its update is still possible, but it requires increased effort and knowledge as it was not designed for it.

Thus, if we need to understand whether risk management can be considered dynamic, we should ask ourselves whether it was intended and designed to be dynamic, as this characteristic must be taken into account in the scope definition, i.e., the very first phases of its development.

2.4 Role of data acquisition in Dynamic Risk Management

Modern data acquisition refers to the process of sampling measurements of a physical phenomenon and converting this into a digital value exploitable by a software. This is usually done using sensors, converters, transmitters, and/or other transmission devices, forwarding the data to an analytical or archiving unit for post-acquisition data processing and/or storing. Manual reporting of information and information extraction from existing databases are usually not considered to be part of the data acquisition process. However, this is generally done in the field of risk management, considering the limited amount of information that can be met when dealing with rare events, as well as the importance of human operators in the management of high-risk infrastructures.

Here, the process of data acquisition can thus be defined as requiring:

- The acknowledgment of the parameter to analyze,
- The identification of the data sources to be considered (measuring device, database, personnel, etc.),
- The capture of the raw data and the initialization of the information transmission.

Although data acquisition may have been considered as part of the risk influencing factor (RIF) "design" in the ORIM [20] or "System feedback" in the BORA [21], it is, to the best of the authors' knowledge, never considered individually as a RIF, and thus never addressed in detail. This is particularly problematic, as the data acquisition process can be failing in different ways. For example:

- data may not be generated (e.g., inactive hardware or hardware failure),
- data may be corrupted (e.g., sensors with false indications),
- considered databases may be outdated,
- data may be available, but not correctly transmitted (e.g., data directed to the wrong endpoint, lack of authority hindering communication in the control team),
- data may be available, but face compatibility issues between devices (e.g., different protocols, language, hardware).

Detecting and acknowledging the occurrence of such issues in real-time is critical to maximizing the probability of good decision making in risk management. Traditional risk management approaches need thus also to be reshaped regarding the possibility of data corruption in order to integrate detection and acknowledgment by design, which is another dimension aimed to be covered by DRM.

3 CONFIRMATION FACTOR FOR VERACITY ASSESSMENT

Expressing the level of knowledge (k) used for risk assessment, as suggested in Formula (2), is an intrinsic feature of the calculated value of risk. We can tolerate having relatively little knowledge of scenarios with both low probability and low consequence. On the other hand, knowledge is critical when the probability and consequences of an event have their highest values. Formula (2) gives important insight into how we should manage risk while continuously improving. As mentioned by several risk management frameworks [14], it should be acknowledged that uncertainty is always a companion [22]. Calibration and correction based on new evidence will possibly allow for decreasing this uncertainty and accounting for evolving system conditions.

Current trends in the IoT and digital fields allow for the assumption that the average number of data sources per parameter observed is likely to increase in the future. Three generic situations may then be encountered when assessing how much the data sources agree on the forwarded information:

1) Absence of alternatives: only one data source is available to inform about a specific parameter. In such a situation, there is no better option than to fully rely on the only existing data source.
2) Confirmation of information: all data sources agree on the information to forward, and the value of the knowledge is increased as the number of data sources increases.
3) Conflicting answers: at least two data sources provide conflicting information. In that situation, further recommendations need to be provided to decide how to handle available information in the risk analysis.

As a consequence, we thus suggest to further characterizing the "strength of knowledge" as reported by Aven and Krohn by splitting the knowledge dimension into two indicators (Formula (3)) to qualify the veracity of information. The first one simply indicates the number (N) of sources available to inform a specific observed parameter. The second indicator corresponds to a veracity indicator (v), capturing the agreement level across the considered data sources, a common approach for data validation and reconciliation [23].

$$k = g(N, v) \quad (3)$$

The number of data sources (N) is reported, as a higher value of (N) would generally imply a higher likelihood that the real status of the observed phenomenon can be captured, especially in the case where the information originates from independent data sources of different nature. The nature of the veracity indicator (v) is based on the nature of the objective function (i.e., categorical, discrete, continuous, etc.), which needs to be clarified in the first phases of the risk analysis. One needs then to define the trust level to assign to each data source, based on a priori knowledge indicating the reliability level of the source. In the absence of relevant side information, an identical trust level will be given by default to the different data sources. Finally, a combination rule needs to be chosen to calculate the value of the veracity indicator.

A simple illustration can be the use of (N) binary data sources (a_i), characterizing the same specific situation (e.g., presence or absence of a hazard in a specified area). Assuming the similar level of trust for all data sources and independence across their acquisition modes (i.e., no common source of corruption), one may choose a simple averaged value to define how likely reported information is to be true:

$$v = \frac{\sum_{i=1}^{N} a_i}{N}, a_i \in \{0,1\} \quad (4)$$

In this situation, the class maximizing the value of (v) may logically be chosen to report the status of the observed phenomenon [24]. Additionally, the value of (v) will enable to better characterize the level of uncertainty existing around the probability dimension reported in Formula (2). However, other situations may be more complex [25] and require a different decision rule, such as counting rules (where at least (M) out of (N) sources need to agree) or linear combinations of the individual data sources, useable when additional information on the reliability of the sources is available. Linear combinations of individual data sources usually outperform simple rules (like the counting rule), while simple rules do not require anything more than monitored information. Finally, discrete scenarios with more than 2 choices and/or fuzzy scenarios where situation evolution may be continuous instead of discrete would also impact the final choice for the decision rule. This decision needs thus to be appreciated on a case-by-case reasoning in the first steps of any risk analysis.

4 CASE-STUDY: VEGETATION MANAGEMENT IN POWER GRIDS

4.1 Context & Data

Overhead power lines are broadly used to transport power from production sites (e.g., dams, nuclear power plants or coal power plants) to consumers (e.g., industrial, commercial, and residential customers). Vegetation represents a main source of hazards worldwide in the management of those power lines [26]. Two principal unwanted scenarios (s) can indeed be identified with this regard: either (1) tree/branch falls on power lines, or (2) vegetation growth under the infrastructures. In both cases, the probability (p) of outages escalates when the distance from vegetation to the power lines decreases, as shortcuts due to connections between different phases are more likely. The consequences (c) can then be particularly important, as this can lead to wildfires and even large blackouts [27]. Distribution System Operators (DSOs) and Transmission System Operators (TSOs) – in charge of the power grid management – require, therefore, to be informed about the presence of vegetation in the surrounding of their grids in order to take adequate maintenance decisions.

Information relative to the assessment of vegetation presence close to power lines is traditionally obtained during visual inspections, which can be executed via foot patrols, helicopters, and the use of drones. Light Detection And Ranging (LiDAR)-based point clouds are also commonly used to obtain 3D insights, allowing for precise distance measurements between power lines and other elements, such as trees. Furthermore, photogrammetry-based point clouds are getting more and more attention in recent years as a more economical alternative to LiDAR point clouds. Finally, the use of orthophotos (geometrically corrected satellite images or large scale aerial images) for efficient large-scale inspections is currently intensively explored [28], mostly pushed by the progress made in computer vision and the continuously increasing availability of satellite imaging technologies with higher resolution and more frequent coverage [29].

These sources of information can all be used to assess the threatening level of vegetation in the surrounding of power grids. Figure 1 (a to d) illustrates how the presence of four small trees growing under a power line can be seen on a drone image (a), in a LiDAR point cloud (b), in a photogrammetry point cloud (c), and on an orthophoto (d).

4.2 Application

The current application principally focuses on the *probability* dimension of the risk definition. We consider the binary case of presence/absence of threatening vegetation between subsections of the power grid (here, a section between 2 consecutive power poles) as a simplified version of the original objective function focusing on exact distance measurement between trees and the infrastructure. In the present situation, the different data sources agree on the presence of four trees growing under the lines, leading the veracity indicator (v) to equal 1. Furthermore, considering the number (4) and the nature of data sources involved, we can confidently assume the information to be accurate. The suggested formulation of the *knowledge* dimension enables thus to dynamically assess the pertinence of the provided probability that threatening trees are present under the lines. It also enables to evaluate the impact of adding/removing data sources in the risk calculation by increasing/decreasing confidence in the provided results depending on the forwarded information.

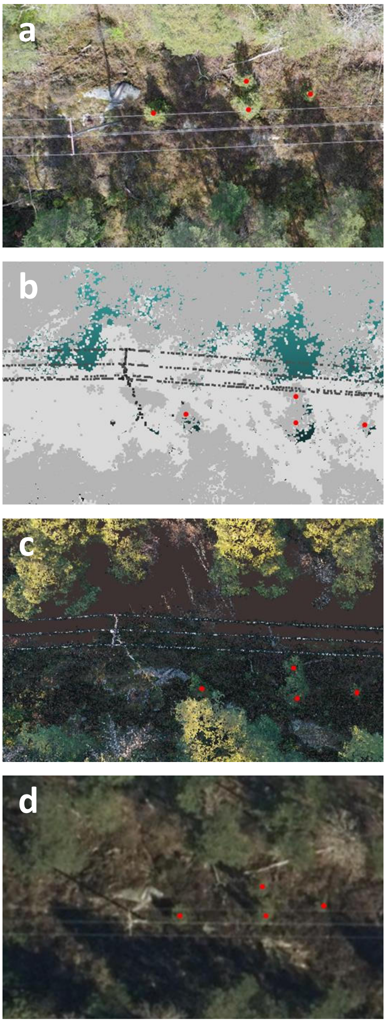


Figure 1 (a to d) - Detection of four small trees (marked via red points) growing under a power line and reported via a drone image (a), a LiDAR point cloud (b), a photogrammetry point cloud (c), and on an orthophoto (d).

5 DISCUSSION

The choice and management of data sources to be considered for risk management remain challenging activities. First, because the pertinence of a choice of data sources may change over time, e.g., due to a variation of the situation circumstances or to a variation in the quality of the acquired data. Second, because the merging of heterogeneous data sources remains a complicated task [30]. For instance, source-specific weaknesses imply that not all data sources can be used for the same purposes. The clarification of the objective function is thus critical to ensure that the chosen data sources properly support the resolution of the problem under review. Furthermore, regular revisions are required to ensure that the choice of data sources is continuously justified. This is, however, hindered by the fact that the definition of the data acquisition process (and more generally, the entire information pipeline construction) is usually seen as a "once and for all" process. This leads it to be often excluded from any serious system improvement plan, even though it has been recognized as a main contributing factor in major accidents. Tracking the variations of the veracity indicator (v) can thus also be used in that sense as a lagging alerting indicator. Once sources of disagreement have been identified, it can potentially reveal the need for removing initially chosen data sources that became irrelevant over time.

The concept of independence across data sources also requires proper attention. Data acquired from independent data sources but in a short period of time may all be outdated when considered in risk analysis, thus all confirming inaccurate information, potentially leading to inadequate decision making. This highlights the critical need for a proper definition of the reliability criteria of the data sources, for which levels need to be tracked over time. Some factors influencing the reliability level of the sources (and thus the pertinence of the executed risk analysis) are, in addition to the choice of the sources, the environmental conditions during data capture, the data collection modes, the choices of technologies, the maintainability of the physical equipment, the exposition of the physical equipment to environmental hazards, the maintenance of the physical equipment and the communication network design. Finally, the choice of the decision rule defining which values will eventually be reported in the risk analyses is a critical task that needs to be adequately executed by a panel of experts in the first phases of the studies and regularly reassessed to ensure its pertinence over time.

6 CONCLUSIVE REMARKS

Ensuring the veracity of information during the entire lifetime of risk management processes is of critical importance to guarantee the pertinence of the reported results. Conventional risk management approaches suffer from a lack of tools ensuring and controlling that data veracity can be kept over time. DRM provides solutions to fill this gap by enabling the establishment of continuity in the risk management processes, facilitating their updates and reiterations when required. This is supported by the rapid development of newly accessible data sources, made, for example, available via numerous IoT-based development strategies. In the present paper, we suggested an approach to formalize the benefit that increased access to a plurality of diverse data sources can provide. For this, we suggested extending the knowledge dimension of a relatively recent risk formulation, based on the number of data sources available and on veracity, indicator capturing the level of agreement across those sources. We applied this approach in a case-study focusing on vegetation close to power lines, which is a common source of outages in power grid management. We concluded that the approach was useful to confirm the presence of potentially problematic vegetation in the analyzed case, but also pointed out that it is, in general, strongly dependent on both the formalization of the problem in initial phases and the quality of the data sources management over the entire life-cycle of the risk management processes.

ACKNOWLEDGMENTS

This work is part of the project "*Dynamic risk management for Smart Grids in large-scale interconnected power systems.*" funded by eSmart Systems and the Norwegian Research Council (NæringsPhD program - Project No.: 276404), which the authors would like to thank for their support.

REFERENCES

1. Thieme CA, Mosleh A, Utne IB, *et al.* Incorporating software failure in risk analysis – Part 1: Software functional failure mode classification. *Reliab. Eng. Syst. Saf.* 2020 ; 197 : 106803.
2. Jacquemain D. *Les accidents de fusion du cœur des réacteurs nucléaires de puissance État des connaissances*. EDP Scienc. 2013 : 464 p.
3. Anderson J, Burkeen AD, Clark D, *et al. Deep Water: The Gulf Oil Disaster And The Future Of Offshore Drilling - Report to the President (BP Oil Spill Commission Report)*. 2011 : 398 p.
4. Ramos MA, Droguett EL, Mosleh A, *et al.* Revisiting past refinery accidents from a human reliability analysis perspective: The BP Texas City and the Chevron Richmond accidents. *Can. J. Chem. Eng.* 2017 ;.
5. Kaplan S, Garrick BJ. On The Quantitative Definition of Risk. *Risk Anal.* 1981 ; 1 : 11–27.
6. Aven T, Renn O. On risk defined as an event where the outcome is uncertain. *J. Risk Res.* 2009 ; 12 : 1–11.
7. Aven T, Renn O. Response to Professor Eugene Rosa's viewpoint to our paper. *J. Risk Res.* 2010 ; 13 : 255–259.
8. Aven T. The risk concept—historical and recent development trends. *Reliab. Eng. Syst. Saf.* 2012 ; 99 : 33–44.
9. IRGC - International Risk Governance Council. *Risk Governance Deficits. An analysis and illustration of the most common deficits in risk governance*. Geneva, 2009 : p.
10. Willis HH. Guiding Resource Allocations Based on Terrorism Risk. *Risk Anal.* 2007 ; 27 : 597–606.
11. Rowe WD. *An Anatomy of Risk*. New York : John Wiley & Sons, 1977 : 24 p.
12. Paltrinieri N, Dechy N, Salzano E, *et al.* Lessons Learned from Toulouse and Buncefield Disasters: From Risk Analysis Failures to the Identification of Atypical Scenarios Through a Better Knowledge Management. *Risk Anal.* 2012 ; 32 .
13. Aven T, Krohn BS. A new perspective on how to understand, assess and manage risk and the unforeseen. *Reliab. Eng. Syst. Saf.* 2014 ; 121 : 1–10.
14. Paltrinieri N, Hauge S, Albrechtsen E. Risk management models in an integrated operations context. *Transactions of the American Nuclear Society*. 2013 : 1854–1857.
15. Canadian Standard Association. *CSA Q850 Risk Management: Guideline for Decision-Makers*. Toronto, Canada, 2009 : p.
16. ISO. Risk Management. ISO 31000:2018. 2018 ;.
17. NORSOK. *Risk and emergency preparedness assessment. Z-013*. Oslo, norway, Norway, 2010 : p.
18. Paltrinieri N, Khan FI. Dynamic risk analysis—Fundamentals. In : Khan FI, Amyotte PRBT-M in CPS, editors. *Advanced Methods of Risk Assessment and Management*. Elsevier, 2020 : 35–60.
19. Yang X, Haugen S, Paltrinieri N. Clarifying the concept of operational risk assessment in the oil and gas industry. *Saf. Sci.* 2017 ;.
20. Øien K. A framework for the establishment of organizational risk indicators. *Reliab. Eng. Syst. Saf.* 2001 ; 74 : 147–167.
21. Aven T, Sklet S, Vinnem JE. Barrier and operational risk analysis of hydrocarbon releases (BORA-Release). Part I. Method description. *J. Hazard. Mater.* 2006 ; 137 : 681–91.
22. Marchi B De, Ravetz JR. Risk management and governance:: a post-normal science approach. *Futures* 1999 ; 31 : 743–757.
23. Narasimhan S, Jordache C. *Data Reconciliation and Gross Error Detection. An Intelligent Use of Process Data*. Houston, Texas: Gulf Publishing Company, 2000 : 411 p.
24. Ciuonzo D, Salvo Rossi P. Distributed detection of a non-cooperative target via generalized locally-optimum approaches. *Inf. Fusion* 2017 ;.
25. Cheng X, Ciuonzo D, Rossi PS, *et al.* Multi-bit Decentralized Detection of a Non-cooperative Moving Target Through a Generalized Rao Test. 2020 ; 1–5.
26. Pacevicius M, A. Ramos M, Paltrinieri N. Optimizing Technology-based Decision-support for management of Infrastructures under risk: The Case of Power Grids. *Proceedings of the 30th ESREL-15th PSAM*. Research Publishing, Singapore, 2020 : 8.
27. Alhelou HH, Hamedani-golshan ME, Njenda TC, *et al.* A Survey on Power System Blackout and Cascading Events Research: Motivations and Challenges. *Energies* 2019 ; 12 : 1–28.
28. European Space Agency (ESA). Space-based Services for Distributed Energy Networks (Smart-Grids). *ESA Bus. Appl.* 2019 ;.
29. Space Industry Bulletin. Market Analysis and Business Intelligence for the Space Community. *Sp. Ind. Bull.* 2019 ; 2 : 24.
30. Pacevicius M, Roverso D, Salvo Rossi P, *et al.* Smart Grids : Challenges of Processing Heterogeneous Data for Risk Assessment. *Probabilistic Safety Assesment and Management - PSAM 14*. Los Angeles, 2018 : 11.

BIOGRAPHIES

Michael Pacevicius
Department of Mechanical and Industrial Engineering
Norwegian University of Science and Technology NTNU
Richard Birkelands vei 2B
7034 Trondheim, Norway

e-mail: michael.f.pacevicius@ntnu.no

Michael Pacevicius is an industrial PhD candidate in the RAMS group at NTNU, Norway and works as a researcher in eSmart Systems, a company delivering high-tech IT solutions for power grid related companies. His research activities focus on the development and implementation of dynamic risk analysis methods for large-scale interconnected power systems. He has a MSc. in Operational safety, Risks and Environment from the Université de Technologie de Troyes (UTT) in France and a MSc. in Economics and Business Administration from the Technische Universität Braunschweig (TUBS) in Germany. He worked as a project coordinator and analyst in the Big Data business development department of SAP in Munich, Germany, before joining eSmart Systems back in 2017.

Nicola Paltrinieri, PhD CEng CSci MIChemE
Department of Mechanical and Industrial Engineering
Norwegian University of Science and Technology NTNU
Richard Birkelands vei 2B
7034 Trondheim, Norway

e-mail: nicola.paltrinieri@ntnu.no

Nicola Paltrinieri is associate professor of risk analysis at NTNU (Norway) and adjunct professor in offshore HSE management at the university of Bologna (Italy). He has earned a PhD in Environmental, Safety and Chemical Engineering from the University of Bologna. From 2012 to 2016 he was research scientist at the department of Safety Research, SINTEF Technology and Society (Norway) and in 2012 he held a postdoctoral position at the university of Bologna. He is chartered engineer in the British Engineering Council register and charted scientist in the British Science Council register. He serves as associate editor of the journal “Safety Science”. He is member of the editorial boards of the “Journal of Marine Science and Engineering”, “Journal of Risk Research” and “Safety in Extreme Environments”. He is member of the board of the NTNU Team Hydrogen. He is member of the board and treasurer of the Society of Risk Analysis – Europe. He serves as Norwegian delegate of the Working Party on Loss Prevention and Safety Promotion within the European Federation of Chemical Engineering. He is co-chair on Accident and Incident modelling, European Safety and Reliability Association Technical Committee. He serves as member of the of the scientific committees for the ESREL, Loss Prevention and CISAP conferences.

Christoph A. Thieme, PhD, MSc, BSc
Department of Marine Technology
Norwegian University of Science and Technology NTNU
Otto Nielsenvei 10
7052 Trondheim, Norway

e-mail: christoph.thieme@ntnu.no

Dr. Christoph Thieme obtained his PhD in Marine Technology from NTNU. He has experience with risk analysis and modelling of autonomous marine systems. Currently, he is a postdoctoral research fellow at NTNU in the UNLOCK project, working on risk assessment methods development and applications on autonomous control systems. His main research interests are the contribution of software and control system aspects to the risk level of autonomous systems.

Pierluigi Salvo Rossi, PhD
Department of Electronic Systems
Norwegian University of Science and Technology NTNU
O.S. Bragstads Plass 2B,
7491 Trondheim, Norway

e-mail: salvorossi@ieee.org

Pierluigi Salvo Rossi received the Dr.Eng. degree (summa cum laude) in telecommunications engineering and the Ph.D. degree in computer engineering from the University of Naples ``Federico II", Italy, in 2002 and 2005, respectively. Currently, he is a full professor of statistical machine learning with the Dept. Electronic Systems, NTNU, Norway, and also the director of IoT@NTNU. He is an IEEE Senior Member since 2011 and serves as an executive editor for the IEEE Communications Letters, an area editor for the IEEE Open Journal of the Communications Society, an associate editor for the IEEE Transactions on Signal and Information Processing over Networks, and an associate editor for the IEEE Transactions on Wireless Communications since 2015. His research interests fall within the areas of communication theory, data fusion, machine learning, and signal processing.

ISBN 978-82-326-5650-9 (printed ver.)
ISBN 978-82-326-6447-4 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology