Knowledge Graph Representation of Power System and Data-Driven Analysis of Outages

by

Yashar Kor

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering

University of Alberta

Abstract

Weather-related power outages in the distribution grid have a significant impact on the grid reliability – they impose a high cost on power utilities and considerable inconvenience to customers. Improvements in monitoring and data collection practices, as well as advanced data processing methods, provide opportunities for comprehensive modeling and analysis of grid operations. At the same time, they allow for better understanding and handling reasons for degradation of service quality due to power outages, weather patterns, and asset-related performance.

The thesis focuses on applying Machine Learning and Computational Intelligence methods for the analysis and processing of power distribution system data. We design and develop a collection of data-driven algorithms, methods, and procedures for investigation of relations between power, outage and weather data. Additionally, they constitute a framework suitable for building a comprehensive system for analyzing and predicting weather-related outages and their severity. We propose *Weather outage Prediction System (WoutPS)* for forecasting outages based on multiple data-driven outage prediction models combined with reasoning framework based on Dempster-Shafer theory (DST), as well as *Knowledge Graph-based representation of distribution grid topology (GridKG)* suitable for integration of data charactrizing different aspects of the distribution system.

Three different architectures of a system for predicting types of weather-related outages are proposed and evaluated. Weather and outage data are utilized for model development and evaluation of their performances. The developed system is capable of identifying the probability of outage occurrences with a focus on identifying outages caused by extreme wind, wet snow, and icing. An analysis of the prediction results is provided.

The thesis includes details of developing a novel knowledge graph based representation of a distribution system. The graph, called *GridKG*, integrates variety of data: system topology, information about its components and customers, as well as data collected during systems events, in particular, power system outages. As a result, a comprehensive representation of a distribution system is obtained. We anticipate that the proposed way of representing a power distribution grid will lead to the discovery of novel ways of augmenting and predicting its reliability. We show the benefits of such representation: evaluation the impact of power outages on consumers in the power system without and with Distributed Energy Resources.

Preface

This thesis is an original work by Yashar Kor. As detailed in the following, parts of the thesis have been published or submitted to journals or conferences in which Prof. Marek Reformat and Prof. Petr Musilek were the supervisory authors and were involved with concept formation and manuscript composition.

Abstract, Chapter 1, and Chapter 2, adopted some sections of the paper Y. Kor, M. Z. Reformat and P. Musilek, "Predicting Weather-related Power Outages in Distribution Grid," 2020 *IEEE Power and Energy Society General Meeting (PESGM)*, 2020, pp. 1-5. A version of Chapter 5 has been published as Y. Kor, M. Z. Reformat and P. Musilek, "Predicting Weather-related Power Outages in Distribution Grid," 2020 *IEEE Power and Energy Society General Meeting (PESGM)*, 2020, pp. 1-5. Chapter 6 adopted some parts of the paper that has been published as Y. Kor, L. Tan, M. Z. Reformat and P. Musilek, "GridKG: Knowledge Graph Representation of Distribution Grid Data," 2020 *IEEE Electric Power and Energy Conference (EPEC)*, 2020, pp. 1-5. Chapter 2, also adopted some sections of this paper. In this paper, as the primary author, I was responsible for concept formatting, designing the algorithms, implementation, and writing the manuscript. Liang Tan assisted with the knowledge graph implementation. Prof. Marek Reformat and Prof. Petr Musilek were the supervisory authors and contributed to the concept formation and manuscript composition. A version of Chapter 6 will be submitted as a journal article.

Acknowledgements

I would like to express my sincere gratitude and appreciation to Prof. Marek Reformat and Prof. Petr Musilek for their unconditional support, ongoing encouragement, and expertise to guide me through this research project. This research and dissertation would not have been possible without their guidance and great supervision. I would like to extend special thanks to my examining committee Prof. Hao Liang for taking the time to review this thesis. I also would like to express my gratitude to my colleagues Antal Buss and Liang Tan for passing on their knowledge and expertise and assisting in my research project. Finally, I am thankful to my family and friends for their unconditional love and support. None of this would have been possible without their encouragement, support and inspiration.

Contents

1 Introduction 1
1.1 Motivation 1
1.2 Objectives and Thesis Outline 2

2 Background 7
2.1 Prediction Models 7
2.1.1 Random Forest 7
2.1.2 Multi Layer Perceptron 7
2.1.3 Support Vector classification 8
2.1.4 Other Classification Models 8
2.2 Definition of Performance Metrics 8
2.3 Basics of Theory of Belief Functions 9
2.3.1 Basic Probability Assignment (bpa) 10
2.3.2 Belief Function 10
2.3.3 Plausibility Function 11
2.3.4 Dempster's Combination Rule 11
2.3.5 Pignistic Transformation 12
2.4 Knowledge Graphs 12

3 Categorization of Power System Protection Challenges 15
3.1 Introduction 15
3.2 Protection Challenge Categorization 17
3.2.1 Distance Relays Challenges 17
3.2.2 Current-Based Relays Challenges 21
3.3 Proposed Protection Schemes in the Literature 23
3.3.1 Adaptive Relays 23
3.3.2 Restricting Wind Farms Fault Current 25
3.3.3 Classification Techniques 26
3.3.4 Modification of Distance Formula 26
3.3.5 Pilot Protection 27
3.3.6 Transient-based Approach for DFIG based Wind Farms 28
3.4 Conclusion 28

4 Outage and Weather Data Analysis 30
4.1 Introduction 30
4.2 Data Description 31
4.2.1 Power Outage and Power System Data 31
4.2.2 Geographic Information System (GIS) Data 33
4.2.3 Weather Data 34
4.3 Integration of Data 34
4.3.1 Outage and System Data 34
4.3.2 Outage, System, and Weather Data 35
4.4 Analysis of Data 36

4.4.1 Outage and System Data 36
4.4.2 Power Outage and Weather Data 45
4.5 Conclusion 53

5 Weather Outage Prediction system – *WoutPS* 55
5.1 Introduction 55
5.2 Related Work 57
5.3 Data Description 59
5.3.1 Outage Data Description 59
5.3.2 Weather Data Description 60
5.3.3 Importance of Parameters for Analysis of Outage . . . 61
5.4 Weather Outage Prediction System – *WoutPS* 61
5.4.1 System Architecture Investigations 63
5.4.2 Data Pre-processing and Model Construction 63
5.5 Outage Prediction Models 65
5.5.1 Model Selection Process 66
5.5.2 Evaluation of *WoutPS* 67
5.5.3 *WoutPS* with Dempster-Shafer Aggregation 70
5.6 Conclusion 71

6 Knowledge Graph Representation of Power System 72
6.1 Introduction 72
6.2 Knowledge Graphs – Brief Overview 73
6.3 Related Work 74
6.4 Power Knowledge Graph 76
6.4.1 Vocabulary Overview 76
6.4.2 Concepts 77
6.4.3 Relations 79
6.4.4 Simple Examples 81
6.5 *GridKG* – Graph-based Representation of Distribution Grid . 82
6.5.1 Electrical Paths 83
6.5.2 Review of Switching Elements 84
6.5.3 Review of System Events 87
6.6 Analysis of Outage Severity 89
6.6.1 Monte Carlo Simulation: Overview 90
6.6.2 Analysis of Simulation Results 93
6.6.3 Analysis of System with Micro-grids 95
6.7 Conclusion 98

7 Conclusions & Future Work 101
7.1 Discussion and Conclusion 101
7.2 Future Work 105
7.2.1 Power System Protection Challenges 105
7.2.2 Outage and Weather Data Analysis 106
7.2.3 Weather Outage Prediction system – *WoutPS* 106
7.2.4 Knowledge Graph for Power System 106

References 107

List of Tables

3.1 Overview of protection challenges and their impact on data-driven power outage prediction models 24

5.1 F-statistic and P-value for weather parameters 62
5.2 Weather parameters importance on the power outage types . . 62
5.3 P-values for student T-test for all model, [144] ©2020 IEEE . 67
5.4 Precision, Recall, and F_1 score report on test data for architectures a, b, c, [144] ©2020 IEEE 69
5.5 Pignistic probability for various samples and the reference values, [144] ©2020 IEEE 70

6.1 Details of 10 outages (selected information) 88
6.2 Mean number of affected customers with 15 power outages in the blue gridcell 99

List of Figures

2.1 Set of few RDF triples with John as their subject 13

3.1 Test system single-line diagram 17

4.1 Schema of the OMS database 35
4.2 Service are grid cell and adverse weather relatd power outages 36
4.3 Weather in rotated pole 37
4.4 Transformers, fuses, and reclosers spatial distribution 37
4.5 Power outage cause frequency 39
4.6 Adverse weather related outages frequency 39
4.7 Protective device frequency (percentage) 40
4.8 Outage duration for various causes 41
4.9 Outage duration for various causes 41
4.10 Frequency of outages for various months 42
4.11 Frequency of outages for various months 42
4.12 Frequency of adverse weather-related outages for various months 42
4.13 Locations of adverse weather-related outages 43
4.14 Clustering of the correlation of equipment and various power outage types 44
4.15 Power outage probability given specific weather condition . . 51

5.1 System Architecture [144] ©2020 IEEE 64
5.2 Cross validation process [145], [144] ©2020 IEEE 66
5.3 Hyper-parameters tuning [145], [144] ©2020 IEEE 67
5.4 Comparison of macro average $F_1, recall, precision$ score of binary classifier (stage 1) used in architecture b, c 68
5.5 Comparison of macro average $F_1, recall, precision$ score of multiclass (stage 2) classifier used in all architectures. 69

6.1 Schema of **PowerKG**: red names represent superclasses of items 80
6.2 RDF triples with *Element-Asset* as their subject 81
6.3 A fragment of the distribution grid with a transformer, three electrical lines and a fuse: electrical connections are shown as thick grey lines. 82
6.4 *Elements* of downstream paths: their types identified by links to light blue circles where each circle represents a different type of electrical component; their voltages – light pink circles; their phases – dark pink circles; as well as outages – red circles. . . 85
6.5 Location of Primary Switch/protective devices on the upstream path from top-left element to the primary breaker (down-right corner). 87
6.6 Details of 10 outage events downstream of a specific *Element-Asset*. 89

6.7 Service area grid cell with power outage locations - blue grid cell represents the cell that power outages in the simulation are located 92
6.8 Number of all affected customers distribution for various number of power outages in the blue grid cell – equipment selection based on: (a) historical power outage probability; (b) uniform probability 94
6.9 Number of all affected customers distribution for $N = 15$ power outages: (a) equipment selection based on historical power outage probability; (b) uniform probability equipment selection. . . 95
6.10 Service area grid cell and customers location - the green grid cell represents the customers that experience power interruption due to power outages in the blue grid cell - the darker colour represents the higher number of affected customers 96
6.11 Number of affected customers distribution in the affected grids (green cells, Figure 6.10) due to power outages in the blue grid – equipment selection based on historical power outage probability. 96
6.12 Number of all affected customers distribution for various number of power outages in the blue grid cell, considering the micro-grids connection to the PCC points: (a) equipment selection based on historical power outage probability; (b) uniform probability equipment 97
6.13 Number of all affected customers distribution for 15 power outages in the blue grid cell considering the micro-grids connection to the PCC points: (a) equipment selection based on historical power outage probability; (b) uniform probability equipment . 98
6.14 Number of affected customers distribution in the affected grids (green cells) due to power outages in the blue grid cell considering the micro-grids connection to the PCC points - equipment selection based on historical power outage 99

List of Abbrevations

AdBo	AdaBoost
AG	Phase-A-to-Ground
ANOVA	Analysis of Variance
BCG	Phase-B-to-Phase-C-to-Ground
CaSPAr	Canadian Surface Prediction Archive
CISG	Converter Interfaced Synchronous Generator
CTI	Coordination Time Interval
CV	Cross-Validation
DER	Distributed Energy Resources
DFIG	Doubly-Fed Induction Generator
DST	Dempster-Shafer Theory
DT	Decision Tree
ECCC	Environment and Climate Change Canada
FACTS	Flexible Ac Transmission System
FRT	Fault Ride Through
GEM	Global Environmental Multiscale
GIS	Geographic Information System
GridKG	Grid Knowledge Graph
HVDC	High Voltage Direct Current
KNN	K-Nearest Neighbor
LAM	Limited Area Configuration
LL	Line to Line
MLP	Multi-Layer Perceptron
NB	Naive Bayes
NoO	NoOutage
NWP	Numerical Weather Predictions
O	Outage
OC	Over Current
OMS	Outage Management System
POTT	Permissive Overreach Transfer Trip
QDA	Quadratic Discriminant Analysis
RDPS	Regional Deterministic Prediction System

RF	Random Forest
RBF	Radial Basis Function
RDF	Resource Description Framework
SAIFI	System Average Interruption Frequency Index
SCIG	Squirrel Cage Induction Generator
SVC	Support Vector Classification
SVCs	Static Var Compensators
STATCOM	Static Synchronous Compensators
SSSC	Static Synchronous Series Compensators
VSC	Voltage Source Converte
WF	Wind Farms
WoutPS	Weather Outage Prediction System

List of Symbols

α	Significance value
β	Relative importance of precision and recall
$Bel()$	Belief function
$BetP_m$	Pignistic probability
$condition_n$	The degree of condition satisfaction in a time window n
$cost(\hat{o}, o)$	Cost of predicting $\hat{o}$ where the true target is o
$confidence_{c_{i,j}}$	Confidence of classifier $c_{i,j}$
f^*	Optimal classifier
$F_{condition}$	Function to determine satisfaction of condition
F_{o_1}	Function to determine outage satisfaction
FN	False negative
FP	False positive
$Fusion()$	Fusion function to combine the results from various classifiers
H_0	Null hypothesis
I^0	Zero sequence current
k	Probability smoothing factor
K^0	Zero sequence compensation factor
$m(A)$	Belief assigned to A
$m_{c_{i,j},ignorance}$	Mass assigned to the ignorance of classifier $c_{i,j}$
N_{equip}	Number of protective equipment in specific cell
N_i^{outage}	Number of outages that equipment i is involved
$\oplus$	Dempster's combination rule
$\mathrm{o}_{(x,y)}$	Random variable representing the number of outages in cell (x, y)
Pr_a	Probability of hypothesis a
$o_1\_condition_n$	The degree of outage and condition satisfaction in a time window n
$Pl()$	Plausibility function
R_f	Fault resistance
s	Machine slip
TP	True positive
Θ	Frame of discernment

ω_i	window weight for hour i
$\mathrm{W}_{(x,y)}$	Multivariate random variable representing weather condition in cell (x, y)
S	Stride
T	Length of time window
Z^+	Positive sequence impedence
ΔZ	Error term in impedance calculation by distance relay

Chapter 1

Introduction

1.1 Motivation

Distribution grid power outages are relatively frequent and impose high costs on power utilities as well as significant inconvenience to customers. According to [1], the US economy is affected by power outages between \$20 billion and \$55 billion annually. Power outages are mainly the result of incidents such as adverse weather, human element, foreign interference, defective equipment, lightning, or tree contacts. These incidents lead to electrical faults in the power system, which should be identified and isolated by power system protective equipment. The protection system aims to keep the power network stable, allowing as much of the network to remain operational as possible while isolating the area under fault. Usually, after power outages, a utility needs to dispatch a large number of crews to restore the interrupted services. The estimated cost of an average storm and the consequent power outages is around \$100,000 to \$1,000,000 per hour [2].

Incidents leading to power outages in the distribution grid result in various types of consequences. Most importantly, it reduces the power system's reliability in providing uninterrupted electricity energy to consumers, as well as it leads to customer dissatisfaction and inconvenience. Reliability in power systems refers to the power system's ability to provide electricity to consumers and satisfy their requirements adequately, and it demonstrates the power system's ability to efficiently responding to system disturbances and resisting uncontrolled events [3], [4]. Furthermore, power outages and utility responses can highly affect the power system's resiliency. Power system resiliency is about grid restoration after power outages and the utility's ability to respond efficiently to limit power outage range, impact, and recover from it quickly [5]. The broader definition of resiliency includes main characteristics such as robustness, resourcefulness, adaptability, and rapid recovery [6], [7]. For reliability analyses, protection systems are presumed to be fully reliable and thus do not account for any malfunctions [8]. The protection system design affects the processes of fault isolation and dealing with power outages.

In recent years, renewable energy technologies for generating electricity have attracted more attention due to climate change and energy sustainability concerns. However, the increasing integration of renewable energy resources poses challenges to the protection system that might lead to its failure. Incorrect functionality of protection systems can cause damages to power equipment, raise safety concerns, reduce power reliability and deteriorate customer satisfaction.

Predicting and determining the severity of weather-related power outages involves two aspects. First, it can be used as a means of identifying locations in the grid that are the most vulnerable to extreme weather conditions, and proposing long-term resilience programs and investments [9]. As a result of these long-term actions, grid reliability, redundancy, and resistance will be improved. Second, it will help with rapid recovery after outages and with the development of emergency plans [10]. Fast recovery plans can prevent the high costs of outages by helping utilities allocate resources in advance of an outage. In addition to increased profitability for utilities, outage predictions improve grid resiliency, reliability, operational efficiencies, as well as customer satisfaction [11].

1.2 Objectives and Thesis Outline

The ultimate goal of this thesis is to analyze power system outages and develop a collection of data-driven algorithms, methods, and procedures that will constitute a framework suitable for building a comprehensive system for predicting weather-related outages and their severity. We believe that prior to power system outage analysis, it is beneficial to understand the interaction between fault incidents and the protection system and be aware of the protection system challenges in the face of the increasing penetration of distributed energy resources (DERs) such as wind farms (WFs).

Categorization of Power System Protection Challenges

Many existing publications on protecting systems with renewable resources have developed their test systems to investigate protection challenges and verify their proposed solutions [12]–[25]. However, the comparison and analysis of the proposed protection schemes studied on different test systems could be difficult.

Therefore, the first objective is to propose a sample ‘test’ system. It allows us to describe the main challenges of protecting power grids with integrated WFs, as well as to discuss the advantages and disadvantages of various relaying algorithms proposed to address these challenges. The goal is to understand the interaction of various WFs with the protection system, and be aware of their influence on the protection relays. Such knowledge would help us estimate how the location and area of power outages will be affected in various configurations. The main objectives are:

- to understand the influence of protection challenges on the consequent power outages and being aware of their impact on the outage location and area;

- to categorize protection challenges and the proposed solutions for various configurations of fault incidents and WFs in the power system.

Chapter 3 provides a description of protection system issues, and responds to the above mentioned objectives.

Outage and Weather Data Analysis

Data collection, integration, and analysis of power outages did not receive enough attention in the literature, and most papers simply used already processed data available for their application.

Our objective is to demonstrate different aspects of the process leading to better understanding of available data. In particular, we aim at:

- providing an overview of the utility outage management system's (OMS) database through the data integration process and demonstrate the relationship between various data sources;

- investigating the interactions between weather, power system, and power outage data and presenting new insights and statistics on various types of power outages; and

- calculating of power outage probability based on weather conditions and similar historical events.

Therefore, in Chapter 4, we focus on data integration processes to construct a unique overview of the utility outage management system (OMS) database. Furthermore, new insights and statistics on the power outages and their relationship to weather conditions are presented. The analysis included in the chapter and its results enable us to better understand the interaction between weather, power system, and power outage data. In addition, to find vulnerable locations to specific weather conditions, we demonstrate the process of calculating the probability of power outages based on similar historical events.

Weather Outage Prediction System – *WoutPS*

As mentioned in the abstract, one of the main objectives of the thesis is to build a weather outage prediction system that has the capability to predict the types of outages. Many papers discuss models for predicting power outages caused by storms such as hurricanes, blizzards, tornadoes, and thunderstorms in the literature [26]–[29]. However, outages caused by wet snow and icing have received little attention in the literature. The goal is to differentiate three major causes of power outages: severe wind, wet snow, and icing, which are the most frequent types of weather-related power outages in Alberta.

We propose data-driven algorithms and methods as a framework for developing a weather-related outage prediction system. Methodologies used to create the framework include Machine Learning algorithms and elements of Computational Intelligence. Furthermore, to provide users with more information to understand the decision-making process involving predictive models, a novel DST-based aggregation method is proposed to provide confidence levels in the prediction. The result is a more powerful ensemble model which aggregates the results of individual classifiers. As a summary, the objectives are as follows:

- construction of a real-time weather outage prediction system capable of predicting major types of weather-related power outages;
- development of reasoning framework to provide confidence in the prediction.

Chapter 5 is fully dedicated to the process of developing of the prediction system. It contains multiple details and descriptions of individual steps.

Knowledge Graph Representation of Power System

The main objective of next chapter, Chapter, 6 is to provide a new representation for a large-scale power system that enables the integration of data sources distributed across various data repositories. This new representation provides an easy and efficient way to grasp a better insight and understanding of the behaviour and mechanisms existing the system via data-driven approaches. The goal is to utilize this representation to investigate the power outage severity impact on utility customers.

Most of the research papers dedicated to the outage analysis discuss predicting how frequent power outages will occur during heavy storms and adverse weather conditions [26], [30]–[32]. However, the number of power outages does not adequately represent the severity of power outages. Yet, there is also a need to estimate the number of customers affected by power outages to determine the actual impact of power outages at different geographical locations.

To address these needs, we develop a power grid knowledge graph (GridKG). It integrates information about the topology of the power system, equipment and customer metadata, and the information about the power outage events in the large scale distribution grid. The proposed GridKG enables deployment of algorithms to identify electrical paths, upstream and downstream connections, and pre-computing the number of connected customers to each piece of equipment in the large-scale grid. Adding these additional data provides a holistic view of the system.

The main objectives of this chapter are as follows:

- to develop a knowledge graph representing a power grid that consolidates various types of data distributed in data repositories and provide a semantically rich representation of power system;
- to design and implement algorithms to enhance the proposed graph with additional information obtained via processing data included in the graph, and to generate more in-depth insight about power outages and their severity.

Therefore, Chapter 6 contains details regarding a process of developing a knowledge graph that provides the ability to integrate very different types of information about the system. The graph will include data about system's topology, its component, type of customers (industrial, residential) and their level of criticality, as well as information about system events – outages, i.e., type of involved equipment, their locations and causes. It will be shown how all this is processed, enhanced and used to estimate the power outage severity.

The final chapter provides the discussion and conclusion, as well as recommendations for future work.

Overall, the thesis is an important step towards advanced analysis of power system outages and utilization of data-driven methods and knowledge graphs in predicting the weather-related power outages and estimating the outage severity impact. It also provides a testimony of the usefulness of such data-driven technologies and a need for continuous data collection in other industrial applications.

Chapter 2

Background

This research is built on several existing approaches and computational methods, briefly outlined in this section.

2.1 Prediction Models

2.1.1 Random Forest

Random forests (RF) is an ensemble learning method that consists of a large number of individual decision trees. Each decision tree (DT) is a tree-like knowledge representation used to classify samples and is constructed in the process of tree building and tree pruning [33]. In RF, each tree is considered a random subset of features and is only trained on a subset of data provided by the bootstrapping method. Individual deep trees can over-fit their training sets, have a low bias, but high variance. However, the random forest model highly boosts the final model performance by averaging individual trees, trained on bootstrapped data, leading to reduced variance at the expense of slightly increased bias.

2.1.2 Multi Layer Perceptron

Multi-layer Perceptron (MLP) is composed of one input layer, one or more hidden layers, and one output layer. Neurons in each layer are connected to the neurons of the next layer and prior layer, parameterized by a weight. The addition of non-linear activation functions to hidden layers and the output layer enables MLP to learn non-linear functions. Weights of MLP are learned through a process of back-propagation, which is a gradient descent on a non-convex objective.

2.1.3 Support Vector classification

Support vector classification (SVC) separates two classes by finding a hyper-plane with the maximum margin between data instances. Originally, SVCs are capable of linear classification; however, by applying a non-linear transformation to the original data, SVCs can be extended to a non-linear separated problem. Obtaining such useful feature representation (kernel) for mapping to a higher dimensional space is a central problem. Common non-linear kernels used with SVCs include polynomial kernel, Gaussian radial basis function (Gaussian RBF), Laplace RBF kernel, hyperbolic tangent kernel, sigmoid kernel. We used the Gaussian RBF kernel since it is one of the most popular and powerful kernels for the non-linear transformation of feature space. In the presence of imbalanced data, SVCs are less likely to have a problem since hyper-plane is learned using a few support vectors. Therefore, the small class size of outage samples may not have a considerable effect on SVCs [33].

2.1.4 Other Classification Models

K-nearest neighbours (KN) classifies the data based on similarity measure and assign each object to the class most common among its nearest neighbours, AdaBoost (AdBo) is an ensemble boosting classifier that combining multiple poorly performing classifiers to make strong classifier, and quadratic discriminant analysis (QDA) enables non-linear separation of data with the quadratic combination of predictor variables [34].

2.2 Definition of Performance Metrics

There are various performance measures that represent the correctness of classification. One of the most popular sets of such measures includes precision, recall, and F_1 score. Precision (P) is defined as the number true positive samples (i.e. correctly classified as belonging to positive class), divided by the number of samples classified as positive, which is the summation of true positive samples and false-positive samples (i.e. incorrectly classified as a positive class). Recall (R) or sensitivity is defined as the number of true positive samples divided by the number of positive samples, which is the summation of true positive samples and false-negative samples (i.e. samples incorrectly not classified as belonging to positive class) [35]. The harmonic mean of precision and recall is defined as F_1 [36].

In order to provide their definitions, we introduce a *confusion matrix* where **O** means Outage, and **NoO** NoOutage:

		Real Value	
		O	**NoO**
Predicted	**O**	$TruePositive : TP$	$FalsePositive : FP$
	NoO	$FalseNegative : FN$	$TrueNegative : TN$

The performance measures are defined in the following way:

$$precision = \frac{TP}{TP + FP} \quad recall = \frac{TP}{TP + FN} \tag{2.1}$$

and

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.2}$$

Generalization of F_1 which gives more flexibility to assigning importance of precision and recall is called F_β and is defined as [37]:

$$F_\beta = \frac{(\beta^2 + 1)\, precision \cdot recall}{\beta^2\, precision + recall} \quad (0 \leq \beta \leq +\infty) \tag{2.3}$$

where β is a parameter that controls the relative importance of precision and recall. If $\beta = 1$, F_β becomes equivalent to F_1. For $\beta > 1$, F_β gives more importance to recall and for $\beta < 1$, it gives more importance to precision such that $F_0 = P$ [36].

2.3 Basics of Theory of Belief Functions

Belief function theory (also known as Dempster-Shafer theory(DST)) has been recognized as a mathematical framework for uncertainty reasoning. It is regarded as an extension of Bayesian probability theory, and it is capable of assigning a mass function to a set of events. Uncertainty is expressed in various mathematical frameworks such as Bayesian probability theory, possibility theory [38], fuzzy set [39], and DST [40]. Bayesian probability is usually the preferred framework when dealing with aleatory uncertainty, i.e., the inherent uncertainty that comes from a random process. This kind of uncertainty comes from the randomness and chance in the underlying variables, which make it irreducible. However, epistemic uncertainty characterized by alternative models is uncertainty in the model of the process, resulting from ignorance and lack of evidence, which is different in nature from aleatory uncertainty. In particular, DST can distinguish the epistemic uncertainty from aleatory uncertainty.

2.3.1 Basic Probability Assignment (bpa)

The finite set of all possible hypothesis or propositions that are collectively exhaustive and mutually exclusive is called frame of discernment $\Theta = \{h_1, h_2, ..., h_n\}$ where the symbol h denotes a hypothesis [41]. Power set of frame of discernment is denoted by 2^Θ and is defined as:

$$2^\Theta = \{A \mid A \subseteq \Theta\} \tag{2.4}$$

A Basic Probability Assignment (bpa) is a function, mapping $m : 2^\Theta \longrightarrow [0, 1]$, which satisfies:

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \subseteq \Theta} m(A) = 1 \tag{2.5}$$

$m(A)$ represents a belief assigned exactly to any $A \subseteq \Theta$, given a piece of evidence. The basic probability assigned exactly to Θ represents the degree of global ignorance, and the basic probability assigned to any subset of Θ which is not a singleton represents the degree of local ignorance [41]. The basic probability assignment will be a classical probability function if there is no global or local ignorance [41].

2.3.2 Belief Function

A belief function, denoted by $Bel : 2^\Theta \longrightarrow [0, 1]$, is defined as:

$$Bel(A) = \sum_{B \subseteq A} m(B), \quad \text{for all} \quad A \subseteq \Theta. \tag{2.6}$$

and satisfies the following conditions:

$$Bel(\emptyset) = 0 \quad \text{and} \quad Bel(\Theta) = 1 \tag{2.7}$$

The belief function represents the total belief in a hypothesis A and its all subsets based on one or more pieces of evidence. Subset A is called the focal set of a belief function, if $m(A) > 0$, and their union is called core. One considerable difference between the theory of belief functions and Bayesian probability theory is that the belief in hypothesis A and the belief in its complement or its negation do not necessarily sum to one. This leads to an additional dimension of uncertainty, which results in discriminating between uncertainty and ignorance and making ignorance explicit [42].

2.3.3 Plausibility Function

A plausibility function, denoted by $Pl : 2^{\Theta} \longrightarrow [0, 1]$, is defined as:

$$Pl(A) = \sum_{B \subseteq \Theta, B \cap A \neq \emptyset} m(B) \tag{2.8}$$

Plausibility refers to the potential belief that can be placed on a hypothesis A if further evidence becomes available [43]. It considers the masses that could be assigned to A and its subsets [41]. Therefore, plausibility can be expressed as $Pl(A) = 1 - Bel(\neg A)$, and it demonstrates the disbelief in the negation of A and shows the degree one fails to doubt in hypothesis A [42].

2.3.4 Dempster's Combination Rule

Dempster's Combination Rule aggregates two or more evidences defined over the same frame of discernment. Let m_1 and m_2 be mass functions defined over Θ, combined mass function $m_{1 \oplus 2}$ is defined as [44]:

$$m_{1\oplus 2}(A) = \frac{\sum\limits_{B\cap C=A} m_1(B)m_2(C)}{1 - \sum\limits_{B\cap C=\emptyset} m_1(B)m_2(C)}, \quad \forall A \subseteq \Theta, A \neq \emptyset \tag{2.9}$$

2.3.5 Pignistic Transformation

Pignistic transformation of a mass function m defined over Θ into a pignistic probability function, denoted by $BetP_m$ is defined as:

$$BetP_m(a) = \sum_{A \ni a} \frac{m(A)}{|A|(1 - m(\emptyset))}, \quad \forall A \in \Theta. \tag{2.10}$$

where $|A|$ is the number of elements of Θ in A [45].

2.4 Knowledge Graphs

Graph-based data formats enable representation of a single entity as a graph node related to other nodes via different relations. Those relations and other nodes can be treated as properties of the entity. Analysis of such data formats can provide insight and better understanding of represented information. It would allow to discover rules that govern relationships between different entities (nodes of a graph), as well as to enable building structures managing synthesis of new information.

The Resource Description Framework (RDF) data format [46] introduced by the Semantic Web as a standard for Linked Open Data is a popular graph-based data format in which each piece of data is stored as a RDF triple. containing two entities, two nodes in a graph, called a subject and an object and a relation between them, an edge in a graph, called a property. Processing data represented as knowledge graphs in an RDF format is generating a lot of attention. There are multiple works focusing on different aspects related to RDF- based Knowledge Graphs: from their construction [47], via storage [48], querying strategies [49], [50] and extracting information [51], to applications [52], [53], just to mention a few. The fact that subjects of triples could be also objects of another triples, and vice versa, means that we deal with a network of entities highly interconnected via properties.

A single RDF-triple <subject-property-object> can be perceived as a feature of an entity identified by the subject. In other words, each single triple is treated as a feature of its subject. Multiple triples with the same subject constitute a description of a given entity. A set of few RDF triples with John as their subject is presented in Fig. 2.1. As it can be seen, each triple provides a piece of description about John: his birth place is Toronto, he lives in Toronto downtown, he likes hockey and swimming, and he works for City of Toronto.

Quite often a subject and object of one triple can be involved in multiple other triples, i.e., they can be objects or subjects of other triples. In such a case, multiple definitions can share features, or some of the features can be centres of other entity descriptions. All interconnected triples constitute a network of interleaving definitions of entities.

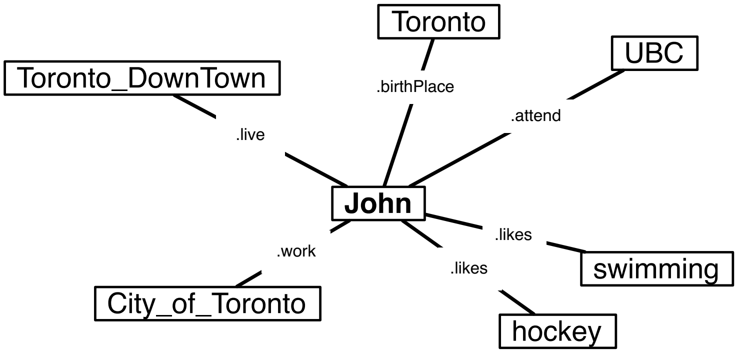


Figure 2.1: Set of few RDF triples with John as their subject

For the purposes of this research, and according to graph database management system, Neo4j, certain terms are defined as follows. *Labels* are used to assign nodes into groups, such as breakers, conductors, switches. Relationships or properties are categorized into different classes by their *types*, such as *Connection* to represent electrical connection for each node. *Uni-directed* graphs have bi-directional relationships, while *directed* graphs have relationships with specific directions. Directions add extra information to the graph; for instance, they can be used to express the energy flow direction. The term *path* is defined as a sequence of nodes and properties which are connected together [54]. For instance, it can be used to demonstrate the sequence of equipment electrically connected together.

Chapter 3

Categorization of Power System Protection Challenges

3.1 Introduction

Renewable energy technologies such as wind, hydro, and solar have attracted more attention due to climate change and energy sustainability concerns. The recent grid-codes require distributed resources to remain connected to the power grid during fault conditions for a specific time in order to increase reliability and access to sustainable energy [55]. As a result, they affect the current characteristics during faults and pose challenges to the protection system. Incorrect functionality of protection systems can cause damages to power equipment, raise safety concerns, reduce power reliability and deteriorate customer satisfaction.

The fault ride-through method used in wind farms can significantly affect their fault current behavior. The literature proposes a variety of methods for improving the fault ride-through capability, including activating crowbar protection circuits [56]–[64], installing flexible ac transmission systems (FACTS) devices, and modifying the converter controller[59], [65]. A converter interfaced synchronous generator (CISG) short circuit behavior depends upon the converter controller, as opposed to a crowbar activated doubly-fed induction generator (DFIG) or squirrel cage induction generator (SCIG) based wind farms that depend upon the induction machine's characteristics. The wind farms should inject positive sequence [66] and both positive and negative sequence currents [66], [67] into the grid during balanced and unbalanced fault, respectively, to meet the recent grid-codes voltage support requirements [66].

The protection system aims to keep power grid stable [68], allowing as much of the grid to remain operational while isolating the area under fault. The protection system of a conventional power system is designed, taking into account that synchronous generators energize the grid. A synchronous generator is modeled as a voltage source in series with an impedance in the fault study of grids. Although such models are helpful for fault analysis and protection design of conventional generators, they cannot be used with systems that incorporate wind farms fault currents because their fault current characteristics differ considerably compared to conventional generators [18].

In general, the grid code requirements, controller structure, deployed fault ride-through method, system topology, intermittent power, and various operating principle of wind farms, in addition to the intrinsic difference between fault response of synchronous and induction generators, are the most critical factors that affect characteristics of fault current and, accordingly, the protection of power grids.

In this chapter we develop a sample 'test' system to provide a better understanding of the protection challenges and their impact on the power grid. We describe the main challenges of protecting power grids with integrated wind farms, and discuss the advantages and disadvantages of various relaying algorithms proposed to address these challenges. Our goal is to be able to understand how various wind farms interact with the protection system and understand how they influence protection relays. Having such information would enable us to estimate how various configurations of wind farms and faults will affect power outages. The main contributions are:

- Development of a sample 'test' system to discuss the main challenges of protection system in the power grids with integrated wind farms.

- Investigation of the various configurations of fault locations and wind farms to address the protection challenges.

- Categorization of the protection challenges and the proposed solutions for power grids with integrated wind farms.

3.2 Protection Challenge Categorization

In this section, we discuss the protection challenges associated with integrating wind farms into power systems. Based on the type of affected relay, these protection challenges can be divided into two categories, i.e., distance-based or current-based.

A sample test system is presented in Fig. 3.1 to describe protection system issues. The test system consists of several candidate locations for wind farm installations and fault locations.

3.2.1 Distance Relays Challenges

Distance relays are commonly used in transmission grids when coordination of current-based relays becomes difficult because of increased time delays. Distance relays estimate the fault distance by measuring the impedance between fault location and relay, which can be expressed by:

$$Z_{distance-relay} = Z^{+} + \Delta Z, \tag{3.1}$$

Positive sequence impedance between fault and relay location, and the error term are represented by Z^{+} and ΔZ, respectively.

The ΔZ error term in a conventional power system results from four key factors, including current injection from adjacent lines, fault resistance [69],

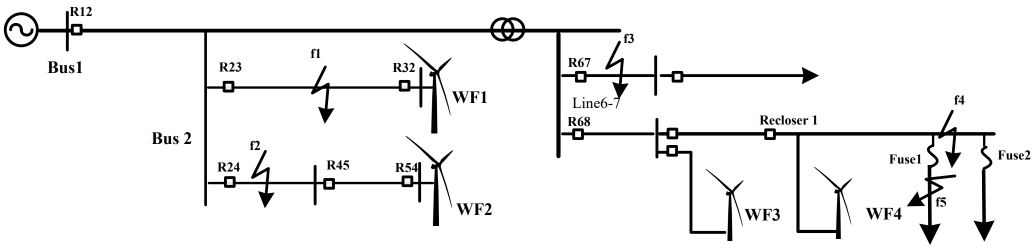


Figure 3.1: Test system single-line diagram

and shunt current injection by compensators [70]–[76]. Adaptive relays are used to address these errors by protection engineers.

Certain wind farm characteristics combined with the sources of error mentioned above result in even greater errors in the estimated impedance in power systems with renewable resources. These characteristics include: wind farm source impedance variations [12], [13], CISG variable phase angle and small fault current magnitude [18], [19], SCIG small fault current after several hundred milliseconds [21], DFIG off-nominal fault current frequency [22], and DFIG and CISG reactive power support [25]. These combinations can result in large inaccuracies in estimated impedances of distance relays and affect the protection system.

Based on the mentioned wind farm characteristics, we can categorize distance relay protection challenges into five groups:

Wind Farm Source Impedance Variations

Variations in wind speed cause fluctuations in wind farms reactive and active power [77], resulting in changes in voltage magnitude ratio and power angle at ends of the line connected to a wind farm [12]. In addition, since the wind farm source impedance depends on the number of generator units connected to the power grid, the impedance can change over time. As a result of these variations, distance relays located on lines connected to wind farms are significantly affected [12], [78]. The impacts of wind farm source impedance variations and FACTS devices on the ΔZ error term are discussed in [14]–[17]. It is shown that the reach settings and measured impedance of distance relays are affected by the type of coupling transformer, FACTS devices, mutual coupling of parallel transmission lines, and the wind farm source impedance variations.

These publications mainly model the wind farm as a voltage source in series with an impedance and consider the pre-fault condition in their protection studies. These models may be used for protection analysis of wind farms based on SCIG and DFIG, but they cannot be applied to CISG wind farms.

SCIG Small Fault Current

After a balanced fault, the current fed by a SCIG drops rapidly, and it can be neglected after 300-400 milliseconds [21], [79]. For a balanced three phase to ground fault at f_2, the impedance error term measured by relay R_{54} can be expressed as:

$$\Delta Z = R_f \frac{I_{54} + I_{24}}{I_{54}} = R_f M_{R_f}, \tag{3.2}$$

According to (3.2), following several hundred milliseconds from the fault instant, ΔZ seen by relay R_{54} for a SCIG based wind farm at WF_2 will be significant as a result of the negligible values in the M_{R_f} denominator. The performance of distance relay secondary zone can be adversely affected by this phenomenon during a high impedance balanced fault; as a result, the relay may not provide adequate backup protection[79]. Despite the failures in the secondary protection zone of the relay, faults in the primary protection zone will be detected accurately due to the large fault current in the transient period [21], [79].

CISG Variable Phase Angle and Small Fault Current Magnitude

Wind farms based on CISG are connected to the power system through a converter. The grid-side converter in CISG controls the voltage on the DC link and supplies voltage support to the power grid, while the generator-side converter adjusts the electromagnetic torque for maximum power extraction from the wind [65]. During a fault condition, for voltage support, the converter current rises to increase reactive power and maintain the voltage at DC link [65]. However, the maximum converter current is restricted by the thermal limits of the converter switches. Thus, CISG fault currents are limited in magnitude while having a variable phase angle. The structure and parameters of the controller determine these fault current characteristics.

In the developed test system, for a fault at f_1 and considering a CISG based wind farm at WF_1 location, the R_{32} relay sees a significant ΔZ factor in its measured impedance due to small fault current contribution from wind farm (local end) and the high fault current from the grid side (remote end).

Additionally, the CISG variable phase angle, which is determined by its controller, leads to a major current phase difference between remote and local end, thereby adding a large imaginary term to the ΔZ factor [20].

Unlike the relay at the wind farm side, the ΔZ error factor for distance relay R_{23}, at the remote end, is relatively small because of the large fault current magnitude in the local end of relay. Therefore, even with variations in the fault current phase angle, the ΔZ factor is small and the measured impedance by the relay R_{23} is more reliable for various types of faults [19].

DFIG Off-Nominal Fault Current Frequency

A DFIG operates similarly to a SCIG after when the crowbar circuit is activated. However, there are substantial differences between the fault currents of the two types of wind farms from a protection point of view. Differences between fault currents result from the range of machine slips. For DFIG, the maximum slip is around thirty percent, as opposed to negligible slip for SCIG [22]. Consequently, the fault current frequency may deviate greatly from its nominal value. Based on the slip value, the fault current frequency can vary from 42-78 Hz for a 60 Hz power system which causes the distance relay's impedance to be estimated incorrectly [22]. For a distance relay at the wind farm terminal, the measured impedance shows a chaotic trajectory during a fault [22]. As a result of such a trajectory, distance relay may fail even if there are no other sources of errors [22]–[24], [80].

DFIG and CISG Reactive Power Support

The wind farms based on DFIG and CISG can inject reactive power through their converters into the power grid, causing voltages to rise during fault events which can adversely affect the distance relays performance [25]. For instance, with a CISG based wind farm at WF_2, and a fault at f_1, the R_{12} distance relay is likely to over-estimate the actual fault distance. This is because the grid-side converter of WF_2 raises the voltage magnitude at bus_2 during the secondary zone time delay of the backup relay. The increased voltage causes a decrease in current on lines connected to bus_2, resulting in a higher impedance estimated by the R_{12} relay. Consequently, the backup relay could overestimate the fault distance, and it will fail to provide reliable backup protection. [25] addresses the problem and presents an impedance calculation method to determine the possible miscoordination in the secondary zone of distance relays.

3.2.2 Current-Based Relays Challenges

In distribution grids, the main protective devices are current-based relays, such as over-current relays, fuses and reclosers. In general, there are three main categories of challenges associated with wind farms and current-based relays, including contribution to fault current, DFIG off-nominal fault current frequency, and its coupled positive and negative sequence responses.

Wind Farms Contribution to Fault Current

The fault current contribution from wind farms during a fault is the major protection problem of current-based relays. We can categorize the impact of wind farms' fault current contributions on the protection system into four groups: loss of sensitivity, loss of coordination, fault current bi-directionality, and recloser failure.

Loss of sensitivity: When wind farms are connected to the grid, they can result in a lower fault current in-feed from the substation, reducing the sensitivity of the main feeder protection devices to downstream faults [81]–[85]. In the test system, the fault current from the wind farm at WF_3 to a downstream fault at f_4, reduces the relay R_{68} reach. Consequently, there can be inadequate backup protection, undetected faults, or delayed fault clearing in case of primary protection failure.

Loss of coordination: An increased fault current may result in loss of coordination between protective devices [81]. For instance the miscoordination of $Recloser_1$ and $Fuse_1$ in Fig. 3.1 is one example. A wind farm located at WF_3 can increase the fault current of a temporary fault at f_5 location. If the fault current exceeds the intersection point between the fast recloser and fuse time-current curves, the fuse will isolate the faulty area instead of the recloser trying to clear the temporary fault [81]. This will cause unnecessary power outages, and data-driven models may underestimate the actual impact of the power outage.

Fualt current bi-directionality: As a result of the connection of wind farms in distribution grids, the fault current can be bi-directional, leading to unnecessary outages in healthy parts of the system [81], [82], [86], [87]. In Fig. 3.1, with WF_3, the relay R_{67} should clear a fault at f_3. However, since the wind farm feeds the fault and R_{68} measures the corresponding fault current, R_{68} could send the trip signal faster than R_{67} which results in the unnecessary disconnection of the $line_{6-7}$ [81].

Recloser failure: In overhead lines, more than 80 percent of faults are temporary. It is possible to extinguish the fault arc and clear the temporary short circuit by disconnecting the fault path from the power grid. Once the electric arc path has been deionized, the line can be reclosed [88]. As a result, the wind farm downstream of the recloser must be quickly disconnected during the reclosing interval to provide enough arc path deionization time [83]. Otherwise, the recloser failure to complete the reclosing operation will result in an unnecessary power outage, and data-driven models may underestimate the actual power outage impact on customers. Fig. 3.1 shows that when WF_4 is connected to the power grid, even during the opening sequence of the $Recloser_1$, it will continuously feed the temporary fault at f_5. As a result, the connection of WF_4 to the grid will prevent the deionization of the fault arc path. Therefore, the recloser will lock out and result in an unnecessary power outage. The data-driven models may underestimate the actual impact of outages in this situation.

In particular, the protection challenges due to the fault current contribution from wind farms, are more significant for SCIG and DFIG. The major reason is that the short circuit level of these wind farms are higher than CISG due to their direct connection to the power system.

DFIG Off-Nominal Fault Current Frequency

The off-nominal frequency of the fault current in DFIG results in the incorrect operation of the directional elements of over-current relays. The current direction in over-current relays is generally identified by measuring the phase difference between current and voltage. Because of the frequency difference between grid voltage and fault current, the phase difference between them changes continuously, resulting in an incorrect direction estimation [89].

Coupled Positive and Negative Sequence Responses

Negative sequence quantities are mainly utilized for ground fault detection in conventional protection schemes [90]. In the presence of DFIG based wind farms, protection schemes based on negative sequence quantities might not be reliable. Without crowbar activation in DFIG, the controller generates negative sequence voltage for reducing the negative sequence current in the stator. In contrast to conventional generators, the DFIG positive and negative sequence responses are not completely decoupled [91], [92], and it differs from a conventional generator. Therefore, the conventional line protection schemes that use negative sequence quantities cannot detect downstream ground faults in grids with DFIG based wind farms. [90].

3.3 Proposed Protection Schemes in the Literature

Various solutions to overcome the protection system challenges have been proposed in the literature. Protection methods can be divided into six categories: adaptive protection, restricting wind farm fault current contribution, classification techniques, distance formula modification, pilot schemes, and transient-based approach for DFIG-based wind farms.

3.3.1 Adaptive Relays

Adaptive protection methods are utilized to address the diverse operating conditions of grid-connected wind farms. According to these methods, relay settings are updated based on the changes in system topology, measured data, and operating conditions of wind farms. These methods are divided into: adaptive distance relays and adaptive current-based relays.

Table 3.1: Overview of protection challenges and their impact on data-driven power outage prediction models

Relay	Characteristic	Likely impact on data-driven models
Distance	wind farm source impedance variations	unreliable protection - overestimation of power outage
	CISG variable phase angle and small fault current magnitud	unreliable protection - overestimation of power outage
	small Fault Current of SCIG	unreliable backup protection - overestimation of power outage
	DFIG off-nominal fault current frequency	unreliable protection - overestimation of power outage
	DFIG and CISG reactive power control support	unreliable backup protection - overestimation of power outage
Current-based	DFIG off-nominal fault current frequency	unreliable direction estimation
	coupled positive and negative sequence responses	unreliable protection - overestimation of power outage
	loss of sensitivity	delayed fault clearing
	loss of coordination	underestimation of power outages
	fault-current bi-directionality	underestimation of power outages
	recloser failure	underestimation of power outages

Adaptive distance relays are designed to handle changes in wind speed, and wind farm source impedance, which can adversely affect the distance relays performance. Distance relays' trip characteristics are automatically tuned by adaptive methods utilizing measurements at the relay location. In addition, to implement adaptive trip boundaries, the number of wind generating units connected to the power grid and local current and voltage measurements at the relay location can be used. Even though such a method performs well, it can only be applied to systems with low levels of wind farm penetration. In power grids with a high integration of wind farms, the impedance and voltage information from the grid-side relay is essential for the relay's proper operation [12]. Based on an artificial neural network, [93] proposed a method to adjust the relay trip boundaries. Although this method reduces the computational complexity, it requires a large amount of memory to handle various possible operating conditions [93]. [14]–[16] develop an adaptive algorithm to generate trip boundaries for distance relays in parallel and single transmission grid with wind farms and FACTS devices. [78] proposes an improved adaptive method for distance relays that uses the pre-fault voltage as reference. The suggested technique dynamically adjusts the reference voltage to compensate for wind power variations,

Wind farms operating condition can significantly impact the fault currents in the power grid, and consequently, a miscoordination may happen if the over-current relays settings are not modified for the power system condition. Several methods have been proposed for adaptive coordination of over-current relays to deal with changing operating conditions of wind farms in the power grid [82], [94], [95]. To address the coordination challenges, [94] uses adaptive protection methods with adaptive group settings calculated through offline analysis for the coordination of the relays. This method can be utilized only when a few wind farms are connected to the power grid; otherwise, many operating scenarios make the coordination task complicated [94]. In the paper, [82] proposes a method for determining the settings of directional over-current relays, including time dialing and pickup currents. The method minimizes the operating time of the primary and backup relay for each setting. An adaptive approach is utilized by [95] to restrict the impact of fault current from converter-based generator such as CISG on the current-based relay operation. This approach uses the grid-side converter controller to adjust the relay settings to compensate for grid topology changes and limit the fault current.

3.3.2 Restricting Wind Farms Fault Current

In order to maintain the relays existing settings like the coordination time interval, several publications propose reducing fault current contribution from wind farms [96]–[98]. According to [96], the converter output current should be adjusted in response to the severity of the voltage drop, i.e., wind farms should reduce the fault current for higher voltage drop. The [97] proposes a method that limits the magnitude of the fault current and, consequently, maintains the setting of current-based relays, taking into consideration the size and maximum capacity of wind generators. The [98] suggests using a superconducting current limiter to restrict fault currents in the power grid with a large wind turbine generator.

3.3.3 Classification Techniques

We discussed in section 3.2.2 that a DFIG with an activated crowbar circuit interferes with the directional elements of current-based relays. As a solution to this problem, [89] proposes a wave shape identification method to identify the direction of current in current-based relays. Unlike conventional generators, the DFIG balanced fault current contains a decaying ac component. Therefore, after the fault instant, the amplitude of the current fundamental frequency component declines over time. Based on this characteristic, [89] introduces an index that can discriminate fault currents from DFIG from those associated with conventional generators. The proposed method is useful when only one side of a protection zone has wind farms; otherwise, the discussed waveshape properties of the current cannot be utilized to determine fault direction.

3.3.4 Modification of Distance Formula

In the presence of CISG based wind farms, the large value of the ΔZ factor results in significant error in the measured impedance of distance relays, as discussed in section 3.2.1. In order to address such large errors, distance relays require modifications to their ground and phase elements. High impedance errors result from the assumption that the phase and ground elements of distance relays are calculated by $Z_{BC} = (V_B - V_C)/(I_B - I_C)$ and $Z_A = V_A/(I_A + K^0 I^0)$, respectively. [20] proposes a distance relay with modified current and voltage signals. The proposed method is able to correctly estimate the fault distance since the modified values reduce the imaginary component of the impedance error [20].

3.3.5 Pilot Protection

A communication link is used in protection systems based on pilot schemes to access the measured voltage and current from each end of the protection zone. In the following several pilot schemes are discussed.

Distance differential method is introduced in [99]. The relay is supposed to have access to voltage and current signals from the remote-end. Then the fault resistance can be calculated by assuming zero line resistance and calculating the active power at each end of the line. Therefore, the ΔZ error term and fault distance can be calculated by knowing M_{R_f}, and R_f values. However, due to assuming a zero resistance for the line, this method may result in an increased impedance error.

For power grids with CISG based wind farms, a pilot scheme with a minimal communication bandwidth is proposed by [20]. It is based on the impedance calculation by a remote distance relay combined with the determination of the current direction by the relays at the local end. Balanced and line to line faults can be identified using this method, while the method in section 3.3.4 is more appropriate for the line to line to ground faults since it does not require communication [20].

[19] proposes pilot protection for the lines with high voltage direct current (HVDC) that is based on voltage source converter technology. As the fault current response of wind farms based on CISG and VSC-HVDC is similar, the proposed method can also protect lines connected to CISG based wind farms. According to the proposed algorithm, faults in the protected zone are distinguished from external faults by the ratio of phase fault currents and negative sequence fault currents at the two ends of the line connected to the wind farm. However, the converter controller of a wind farm can inject both positive and negative sequence currents into the power grid during unbalanced faults, which may jeopardize its performance.

3.3.6 Transient-based Approach for DFIG based Wind Farms

In DFIG, the crowbar activation causes an error in conventional impedance estimation methods that are based on the nominal frequency component of current and voltages. [23] proposes a transient differential equation-based distance protection scheme along with a faulted phase selection algorithm to minimize the impact of the off-nominal frequency component of the fault current on the fault distance estimation. Because this scheme is a transient-based approach, it functions properly before and after the activation of the crowbar circuit. For ground relaying of lines emanating from DFIG based wind farms, the zero-sequence current is used rather than the negative sequence current [90].

3.4 Conclusion

In this chapter, we investigated the power system protection challenges in the face of increasing penetration of wind farms into the power grid. We developed a sample 'test' system for the comparison of different protection algorithms. Additionally, we discussed various reasons and configurations that may lead to the malfunction of the protection system. We investigated several solutions proposed in the literature to address the protection-related challenges. The proposed protection schemes associated with current-based relays include restricting wind farm fault current, classification techniques, and adaptive methods. It is possible to use the existing relay settings by limiting the converter output current in CISG and restricting the maximum fault current using superconducting fault current limiters. However, superconducting fault current limiters increase the system cost and require proper protection coordination. Furthermore, restricting wind farm fault current schemes do not address the challenges associated with the diverse operating conditions of wind farms. We further discussed that the protection systems based on fault classification techniques could be used when wind farms are located at one side of the protection relay. We suggest that adaptive current-based relays, which automatically adjust the relay settings based on the measurements at the relay location and system topology, are the most effective approach for addressing the challenges. Furthermore, we discussed that adaptive distance relays could be used to address the challenges related to wind farm source impedance variations. In distance relays, pilot relaying algorithms appear to be the most effective protection scheme. In conclusion, we believe that understanding the problems that the power system may experience has significant importance for the power utilities to prepare themselves while integrating the grid with renewable energy resources such as wind farms. If not addressed in advance, these challenges can affect the expected behavior of the protection system, as well as the location, severity, and impact of power outages.

Chapter 4

Outage and Weather Data Analysis

4.1 Introduction

The process of data analysis and converting them into future insights requires access to a sufficient amount of high-quality data. Data collection, integration, and pre-processing are essential to ensuring high-quality data is used in the research studies. Therefore, in this research, great effort is put into the data processing and explaining them.

Furthermore, this chapter analyzes the prepared data sets and summarizes the main characteristics of the available data using statistical graphics and data visualization. The goal is to understand what the data can tell us ahead of the formal modeling. This process is essential for understanding the interactions among the variables, which could suggest hypotheses regarding the observed events and their causes. Moreover, finding the important variables for power outage analysis and formulating the probabilistic model for power outages were another objectives of this chapter.

In the literature, the data collection, integration, and analysis of power outages did not get enough attention, and most of the research papers used the previously available data. However, we believe it is essential to demonstrate the process to better understand the discussed variables and help to conduct similar research in other industrial applications. The main contributions of the chapter are:

- providing an overview of the utility outage management system's (OMS) database, through the data integration process, and demonstrating the relationship between various data sources;
- presenting new insights and statistics on various types of power outages;
- demonstrating the interaction between weather, power system, and power outage data by transforming power outage data at the grid cells level and integrating it with weather data;
- formulating the predictive inferences based on posterior predictive distribution of the power outages and the weather condition.

4.2 Data Description

The main data sets that include power outage, system data and weather data are utilized in the thesis.

4.2.1 Power Outage and Power System Data

At most power utilities, the power outage data are stored in relational databases called outage management systems (OMSs). The power outage data utilized in this dissertation is collected by the OMS system of a major energy holding company in Alberta, Canada. Power outage data are essential for recognition causes of outages, their locations, and their effects. In this section, we focus on integrating, storing, and accessing data. A variety of data and their formats makes this task an important and necessary one for analyzing power outages.

The information regarding the customers' calls and reporting outages are stored in the "outcall" table. When each customer calls and reports an outage, a new row with a unique "event_id" is created to store the information such as the outage id, customer location, time of the call, customer transformer number, and customer account and its premise number. The information from the customer call starts the process of outage restoration if it is not a scheduled power outage. The affected area that experiences power interruption is considered downstream of the location of the first upstream protective device. When several customers report an outage, the outage location is updated considering the location of each customer, which may imply a greater affected area.

The information regarding outages is stored in the “aeven” table. For each outage, a unique “outage_id” is assigned to identify the outage. Furthermore, the device that cleared the fault, i.e., “device_id”, outage restoration time, substation, and feeder number, are also included in this table.

Power outage causes and their description are stored in a separate table “outage_complete”. The primary key in this table is “event_id,” and the outage cause descriptions are provided as primary and secondary causes of the power outages. Primary causes of outages include:

- Loss of Supply: Power outage due to interruptions of power from the bulk electricity system (transmission grid) resulting from problems such as maintenance on the transmission grid, under-frequency load shedding, transmission system transients, and system frequency excursions.

- Tree Contacts: Power outage due to faults caused by contacting trees or tree limbs to the energized section of the grid.

- Lightning: Power outage due to lightning striking the grid and resulting in flashovers or insulator breakdowns.

- Defective Equipment: Power outage due to failure of defective equipment because of lack of maintenance or age.

- Adverse Weather: Power outage due to extreme weather conditions and as a result of extreme wind, icing, snow, ice storms, and other adverse conditions.

- Adverse Environment: Power outage due to abnormal environments such as fire, corrosion, humidity, salt spray, and flooding.

- Human Element: Power outage due to deliberate damage or utility staff errors and incorrect settings, installation, and protection settings.

- Foreign Interference: Power outage due to out-of-control events such as animals, vehicles, birds, and foreign objects.

- Scheduled Outage: Power outage due to disconnection of part of a power system for maintenance and construction.
- Unknown: Power outage due to no apparent cause or reason for the power outage.

Furthermore, detailed secondary causes are provided to add more information regarding each outage. The primary and secondary causes are represented as a “prime_cause”, “sup1_cause”, and “sup2_cause”, which describes each outage’s main reason and supplementary information.

The list of transformers with power interruption for each outage is provided in the “outhist_transformers” table. For each transformer, information such as the number of connected customers and the duration of outages is provided.

The customers’ information is stored in the “cispersl” table. It includes the customers’ unique id, account number, premise number, and the transformer id, to which the customer is connected.

The information regarding the connection of each electrical equipment is provided in the “oms_connectivity” table. This table provides information such as the geographical coordination of each piece of equipment, type of device, last modification date, and two electrical node values. Each piece of equipment has two nodes, and if two pieces of equipment are connected, they share the same node value. Thus, having the equipment node values, the topology of the system, and their connections can be extracted from this table.

4.2.2 Geographic Information System (GIS) Data

GIS data provides information about each specific type of equipment in the electric grid. The GIS format includes vectors to represent spatially referenced data and attribute tables for tabular information. Vectors are classified into polygon, lines, and point data and are utilized to represent regions, conductors, and other equipment, respectively. The tabular data provides the technical information regarding each piece of equipment. The topology of the power system can be inferred both from the “oms_connectivity” table and GIS data. It should be noted that GIS provides more accurate geographical information for each piece of equipment. For, instance the exact location and shape of conductors are available in GIS data, while the “oms_connectivity” table only provides the approximate location of the center of the conductor and its node values.

4.2.3 Weather Data

Weather data are required for learning characteristics of extreme conditions, such as wind/winter storms. To obtain that data, we used the Canadian surface prediction archive (CaSPAr). Weather data preparation involves the collection and storage of data for offline processing. This task focuses on two activities: 1) collecting weather data representing weather conditions and 2) preparing weather data as inputs for analysis.

Weather data in NetCDF4 format is collected from the CasPAr platform, which provided the numerical weather predictions (NWP) archive issued by Environment and Climate Change Canada (ECCC). For this study, we used the regional deterministic prediction system (RDPS), which provides hourly simulation of weather parameters with a rotated longitude-latitude grid with $\sim$ 10 km resolution. The RDPS is based on limited area configuration (LAM) for the Global Environmental Multiscale (GEM) model, which covers North America [100].

Wind-related variables are calculated at 10 m, temperature and due point values are considered at 1.5 m, and the rest of the variables are considered at the surface level.

4.3 Integration of Data

4.3.1 Outage and System Data

The data from the utility OMS is used to extract the information regarding distribution grid power outages from 2015 to 2019 in Alberta, Canada. The outage start time is approximated when the first customer reports the outage to the utility and is found out from the “outcall” table. Outage location is approximated, the location of the first upstream protective device, i.e., overcurrent relays, switches, reclosers, fuses, and transformers from the outage location report. The outage location is updated after the report of new customers. The protective equipment location is extracted from the "oms_connectivity" table, which stores all the electrical equipment locations, node values, and ids. Nested outages are distinguished as they have the same start time but different restoration times, and they are considered unique outages. Outage types are categorized into nine groups based on the information provided in the "outage_complete" table. For each outage, the list of affected customers is extracted by cross-referencing the "aeven", "outhist_transformers", and "cispersl" tables. Furthermore, the "oms_nonconnect" table that provides the geographic coordinates of customers and non-electrical equipment is used to determine the geographical coordination of customers.

The schema of the database and one sample of outage data is depicted in Figure 4.1.

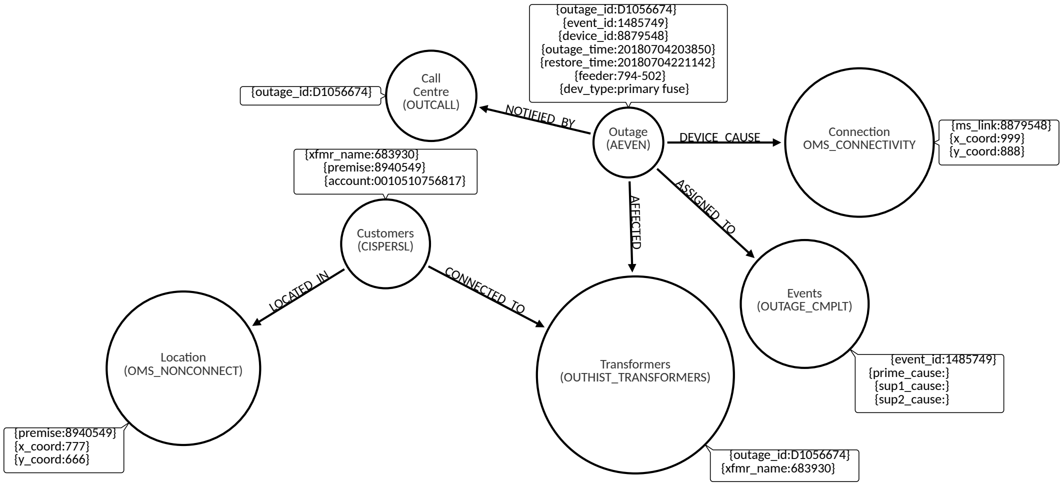


Figure 4.1: Schema of the OMS database

4.3.2 Outage, System, and Weather Data

In order to find out the weather condition for power outages, the nearest weather data point to the outage location are considered. A grid-cell based on the ∼10 km weather grid is defined, and a mesh of grid cells with weather data points at the center of each grid cell is created. The weather condition is considered to be the same across each grid cell, and the power system and outage data are aggregated at the ~ 10 × 10 km grid cells. Figure 4.2 shows the created grid cells for one service area in Alberta, which demonstrates the grid cells, their centre, and the location of various weather-related power outages.

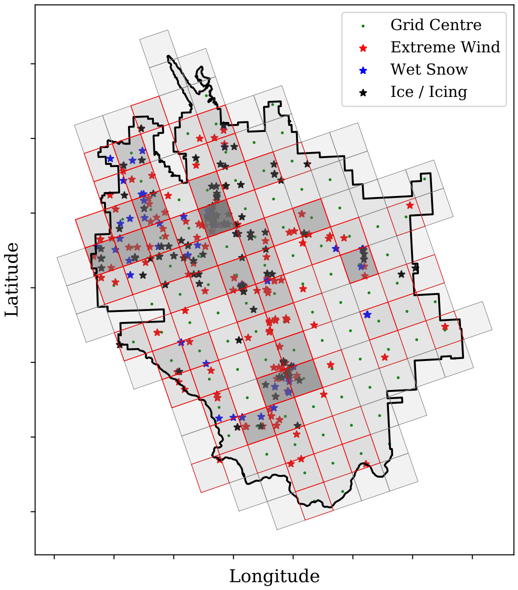


Figure 4.2: Service are grid cell and adverse weather relatd power outages

The weather data are stored on the native rotated latitude and longitude grids issued by ECCC. For faster access to the weather data, the power outage longitude and latitude are transformed to rotated longitude and latitude, and then the corresponding weather data are queried from the weather database. The grid cells in the rotated grid and the plot for temperature across the grids are depicted in Figure 4.3.

The equipment data are aggregated at the grid cell level, i.e., the number of each type of equipment is calculated for each grid cell. Figure 4.4 demonstrates the location of transformers, fuses, and reclosers and the corresponding grid cell with the color representing the density of each type of equipment.

4.4 Analysis of Data

4.4.1 Outage and System Data

Some analyses about power outage data are provided in this section to better understand the characteristics of the studied power outages.

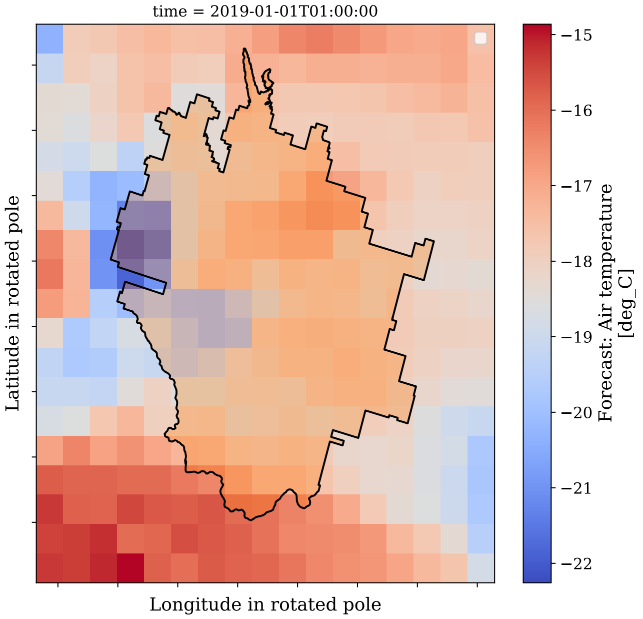

time = 2019-01-01T01:00:00
Latitude in rotated pole
Longitude in rotated pole
Forecast: Air temperature [deg_C]
−15
−16
−17
−18
−19
−20
−21
−22

Figure 4.3: Weather in rotated pole

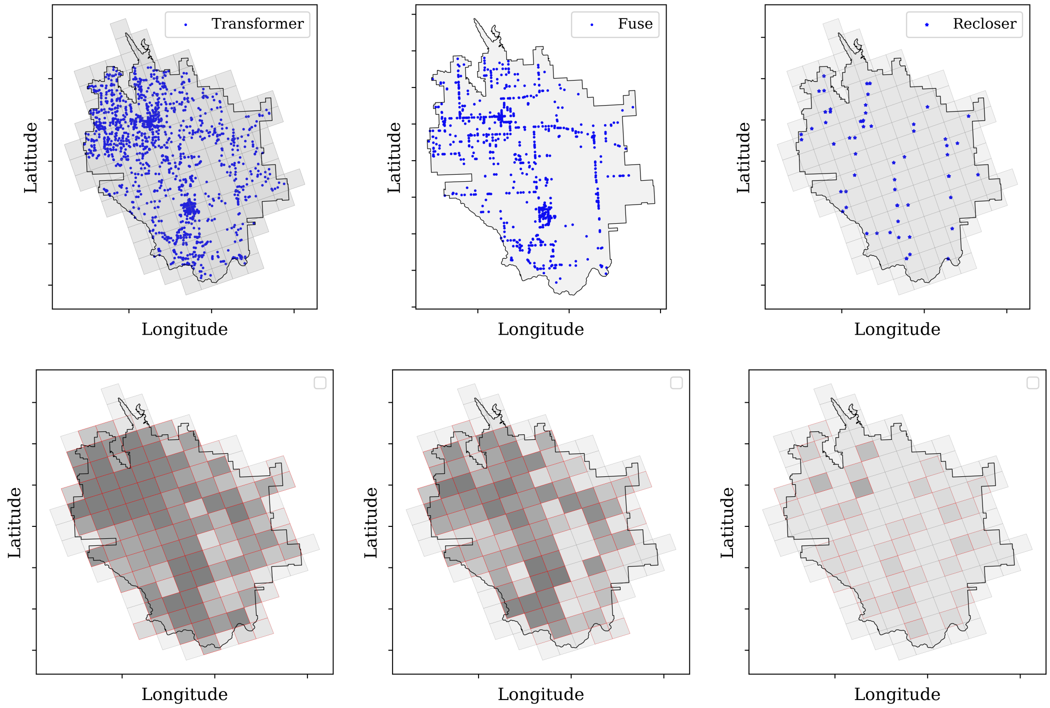


Figure 4.4: Transformers, fuses, and reclosers spatial distribution

Primary cause of outages

The primary cause of power outages and their frequency from 2015 to 2019 in Alberta, Canada is depicted in Figure 4.5. The loss of supply or interruptions of power from the transmission grid is only responsible for 0.5% of outages, which indicates that most of the power interruptions that customers experience are due to the incidents in the distribution grid. The most frequent cause for interruption of power is scheduled outages, mainly due to maintenance and construction. The distribution grid consists of numerous equipment and thousands of kilometres of power lines that justify periodic maintenance and construction. The second most frequent reason is reported as unknown outages, which means the utility could not find any apparent reason for the outage. Power utilities usually send crews to find the outage cause and restore the power after the incident is reported.

The power outage locations are mainly approximated by the first upstream protective device, which clears the fault and isolates the downstream grid, and it can be far from the location of the incident that triggered the outage. Therefore, it can be very costly for utilities to search for the exact location and find the outage cause if the incident did not cause permanent damage to the power grid.

Power outages due to defective equipment are mainly because of electrical or mechanical failure in equipment such as fuses, pole mount transformers, primary conductors, secondary cables, tie wires, or pin insulators.

Foreign interference, which represents more than 11% of power outages, is mainly due to out-of-control events such as wildlife (birds, animals) which mainly cause phase-to-phase and phase-to-ground short circuit faults. Furthermore, incidents caused by vehicles, agricultural equipment, digging in, and vandalism are the next important causes of foreign interference outages. Adverse weather is responsible for more than 9% of power outages in the distribution grid.

Adverse weather and tree contacts and lightning are the major reasons for unscheduled power outages. The main reasons that adverse weather conditions lead to power outages are extreme wind, wet snow, icing, and freezing rain. Among adverse weather-related power outages, icing, extreme wind, and wet snow are almost equally frequent in the distribution grid.

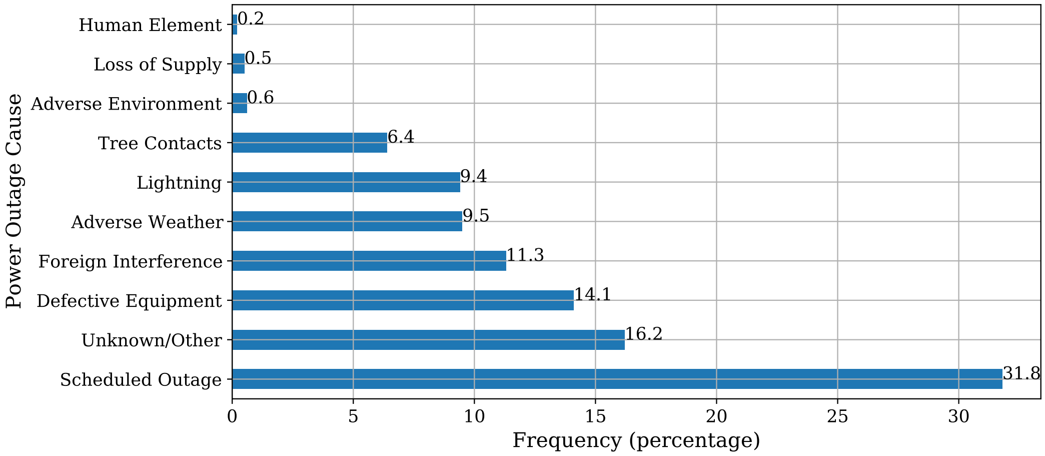


Figure 4.5: Power outage cause frequency

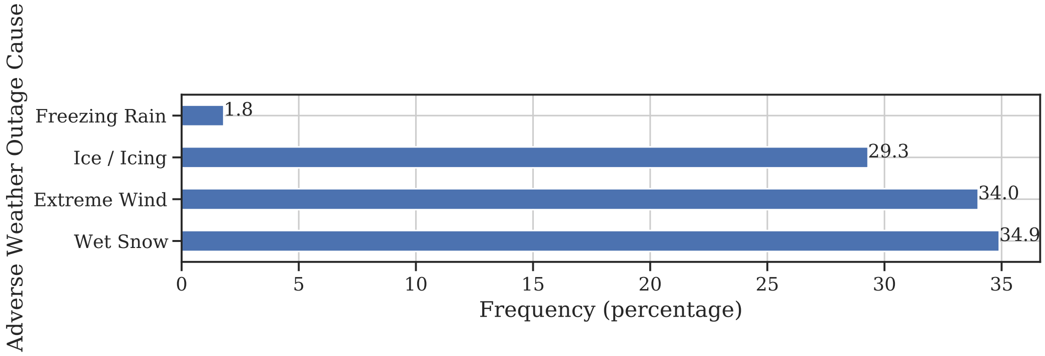


Figure 4.6: Adverse weather related outages frequency

Protective device

The protective devices are responsible for clearing the faults and minimizing the affected area by isolating part of the power grid. As depicted in Figure 4.7, line and transformer fuse clear the faults in more than 85% of the power outages. It demonstrates the importance of fuses in the protection of the distribution grid. When fuses blow, the downstream grid is disconnected permanently, and the utility should send the crews to restore the power. Therefore, to minimize power outage duration for temporary faults, reclosers are used to disconnect the grid for a short time and restore the power if the fault is cleared. However, in case the fault is permanent, the recloser disconnects the downstream grid. As shown in Figure 4.7, reclosers cleared the permanent faults in the distribution grid in more than 8% of power outages. Switches such as air break, solid, and ganged air break are the main types of switches utilized in the distribution grid and cleared 3.8% of the faults. Primary circuit breakers located at the beginning of the feeders are the primary protective devices that disconnect the whole feeder. Although the primary breakers clear only 1.2% of the faults, they cause a power outage for all the customers connected to that feeder and affect more customers.

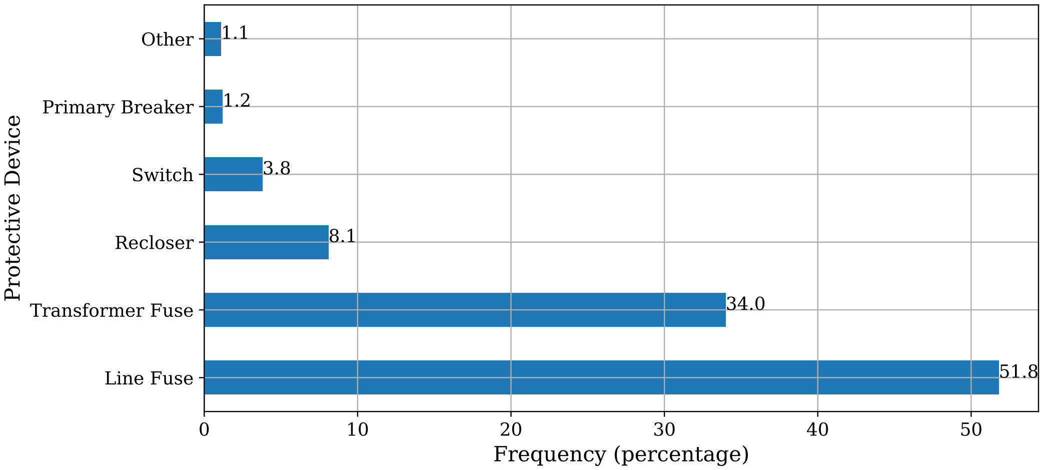


Figure 4.7: Protective device frequency (percentage)

Duration of outages

The bars in Figure 4.8 show the mean value of outage duration for each outage cause. The error bars show the uncertainty around that mean value, and they indicate the 95% confidence interval for the population mean value. The adverse environment has the highest outage restoration duration. Events such as fires or floods usually affect more areas, and they need more time and effort to restore the power. Outages due to adverse weather also have a relatively long duration compared to other causes such as scheduled outages. Figure 4.9 depicts outage duration for secondary causes of adverse weather conditions, including extreme wind, ice/icing, freezing rain, and wet snow power outages.

It shows that restoring power outages due to wet snow takes more time than other weather-related power outages.

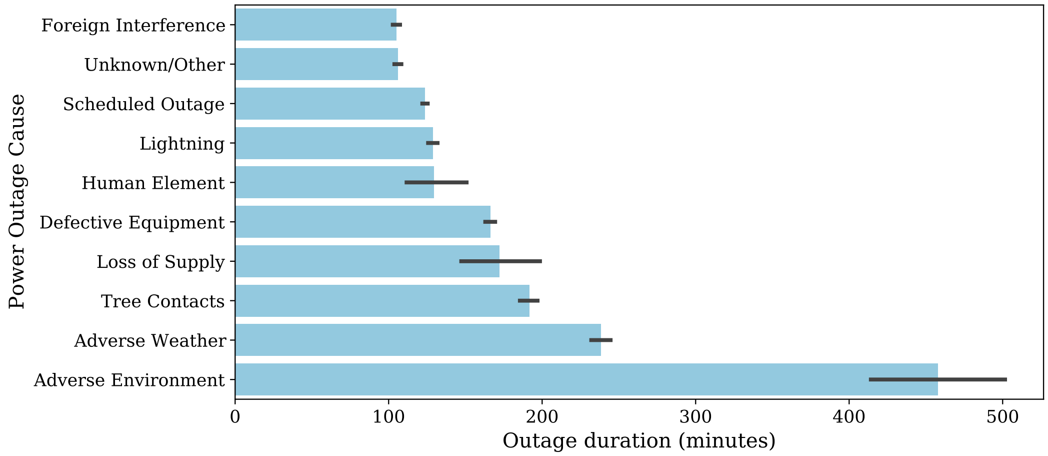


Figure 4.8: Outage duration for various causes

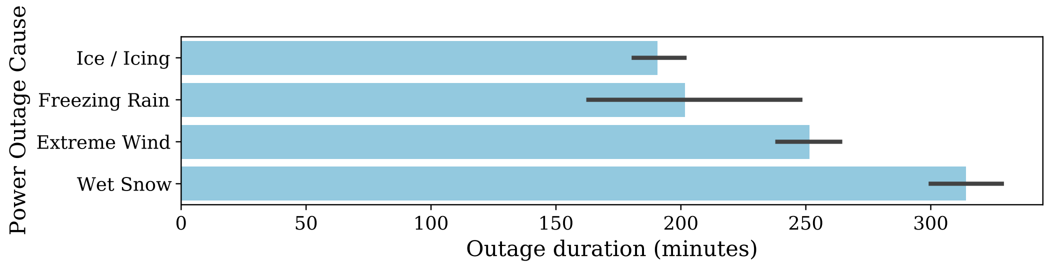


Figure 4.9: Outage duration for various causes

Month

Figures 4.10, and 4.11 show the frequency of power outages for each month and outage cause. It can be seen that tree contacts, lightning, foreign interference, and adverse environment outages mainly happen in the summer months in Alberta due to a higher number of fire events, lightning, tree growth, and animal activities during that time. The number of scheduled outages is lower during December, January, and February, but it is almost the same for the rest of the year. Figure 4.12 depicts the frequency of adverse weather-related outages for various months. Outages due to extreme wind mostly happened in June, mainly because of tree growth. Outages due to wet snow mainly occurred in April, May, and October when temperatures are around zero, while Icing outages mostly occurred during cold weather conditions in October, November, and January.

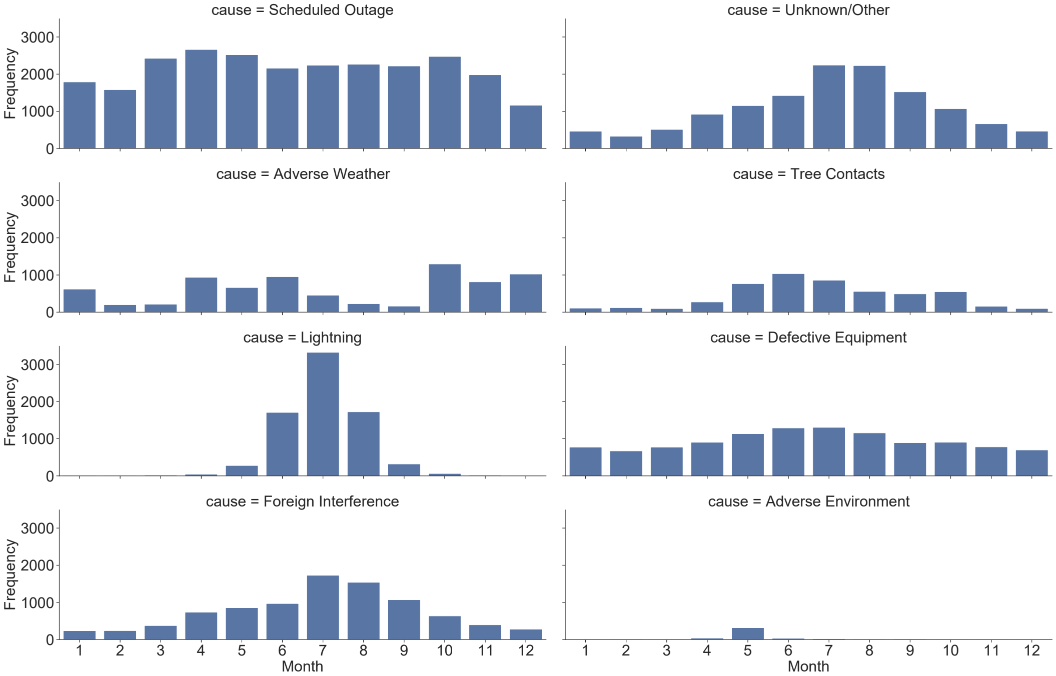


Figure 4.10: Frequency of outages for various months

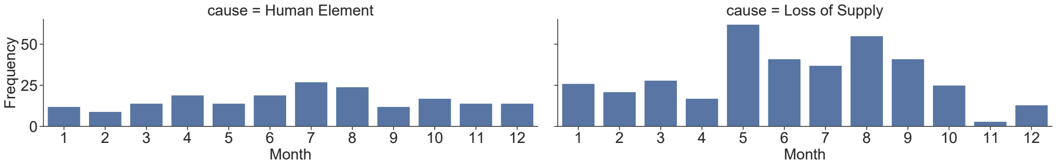


Figure 4.11: Frequency of outages for various months

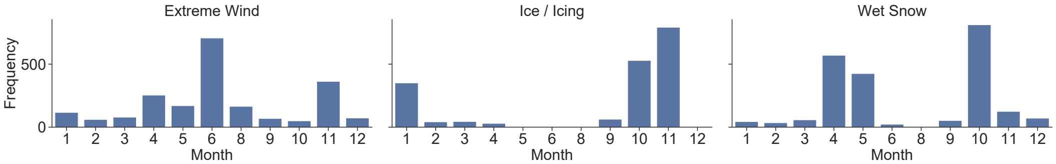


Figure 4.12: Frequency of adverse weather-related outages for various months

Locations

Figure 4.13 depicts the location of outages due to adverse weather, including extreme wind, wet snow, and icing conditions. It can be observed that location has a considerable effect on the type of power outages. Some locations are more vulnerable to specific types of outages due to equipment, vegetation, and geographical features.

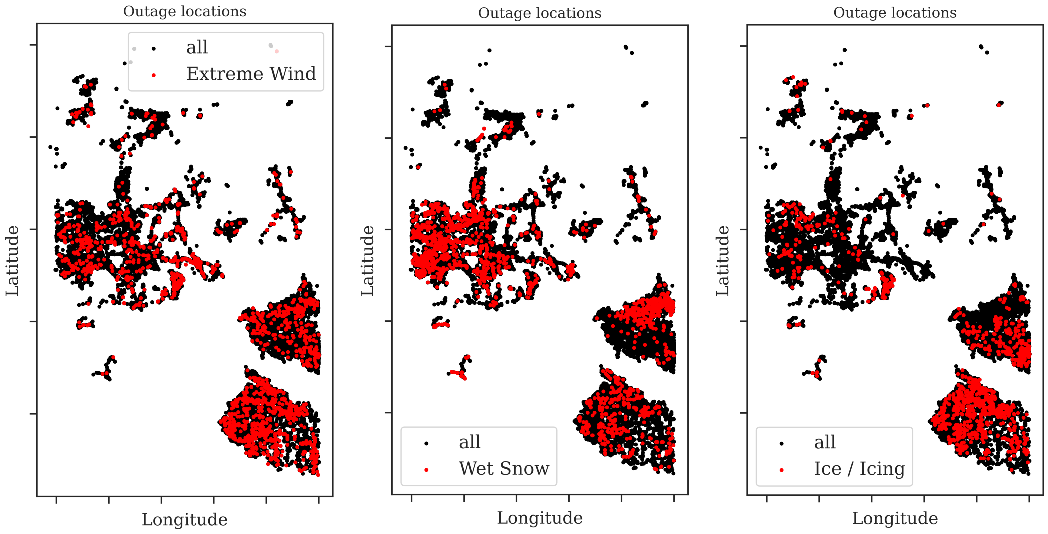
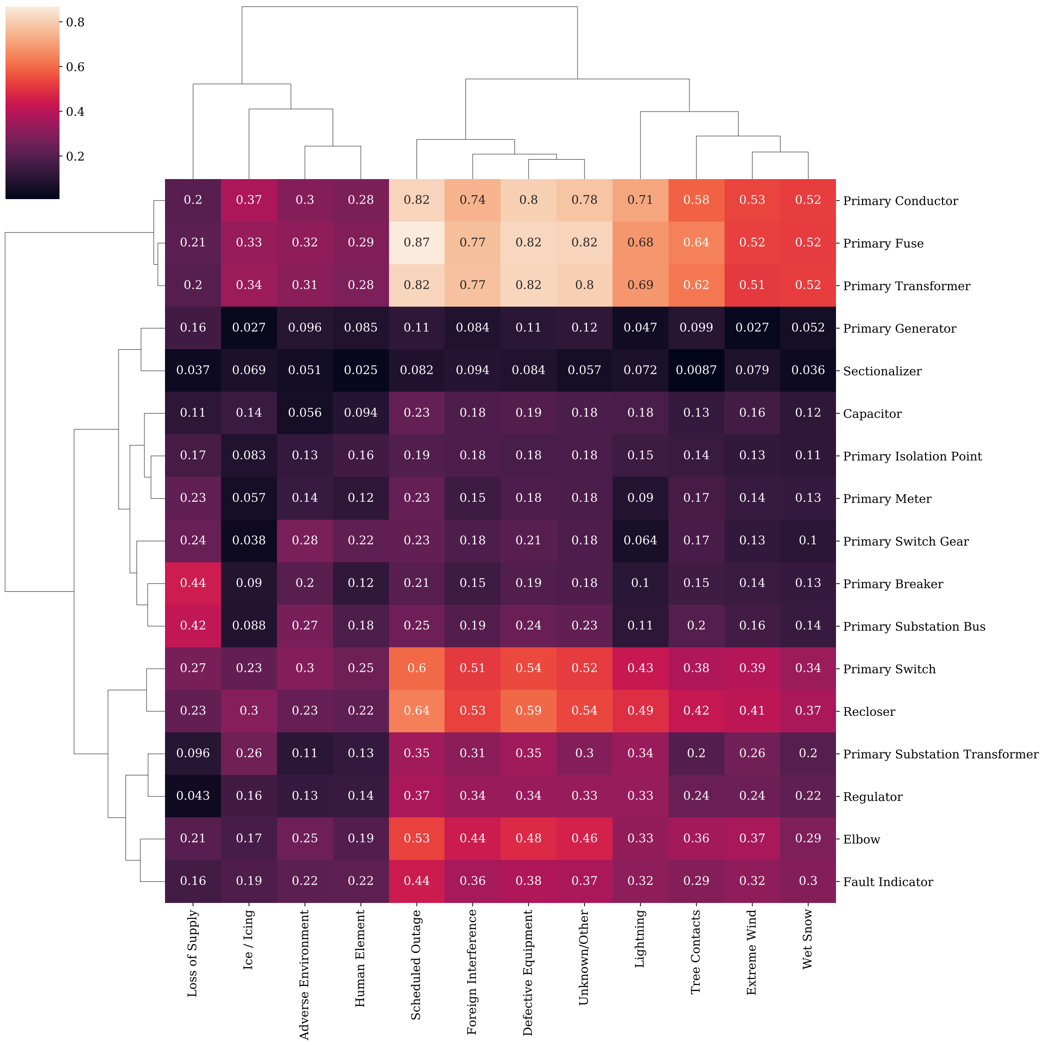


Figure 4.13: Locations of adverse weather-related outages

Correlation between power outage and power equipment

In this section we are interested in understanding the relationships between frequency of various types of power outages and the number different equipment in each area. There is not evidence that the number of power outages and the number of equipment are normally distributed across the grid cells. Therefore, non-parametric methods are used to determine whether there is a statistical relationship between variables. Spearman rank correlation is calculated as a measure of monotonic association between the number of equipment and various outage types. This test determines if a monotonic function can be used to describe the relationship between the variables. The correlation coefficient is scaled between 1 and -1. The value around 0 means there is no monotonic associations between variables, and the values approaching +-1 indicate a strong increasing or decreasing monotonic relationship. Figure 4.14 demonstrates the calculated correlation coefficients based on the information of more than 800,000 equipment and 100,000 power outages across all grid cells in the province.



	Loss of Supply	Ice / Icing	Adverse Environment	Human Element	Scheduled Outage	Foreign Interference	Defective Equipment	Unknown/Other	Lightning	Tree Contacts	Extreme Wind	Wet Snow
Primary Conductor	0.2	0.37	0.3	0.28	0.82	0.74	0.8	0.78	0.71	0.58	0.53	0.52
Primary Fuse	0.21	0.33	0.32	0.29	0.87	0.77	0.82	0.82	0.68	0.64	0.52	0.52
Primary Transformer	0.2	0.34	0.31	0.28	0.82	0.77	0.82	0.8	0.69	0.62	0.51	0.52
Primary Generator	0.16	0.027	0.096	0.085	0.11	0.084	0.11	0.12	0.047	0.099	0.027	0.052
Sectionalizer	0.037	0.069	0.051	0.025	0.082	0.094	0.084	0.057	0.072	0.0087	0.079	0.036
Capacitor	0.11	0.14	0.056	0.094	0.23	0.18	0.19	0.18	0.18	0.13	0.16	0.12
Primary Isolation Point	0.17	0.083	0.13	0.16	0.19	0.18	0.18	0.18	0.15	0.14	0.13	0.11
Primary Meter	0.23	0.057	0.14	0.12	0.23	0.15	0.18	0.18	0.09	0.17	0.14	0.13
Primary Switch Gear	0.24	0.038	0.28	0.22	0.23	0.18	0.21	0.18	0.064	0.17	0.13	0.1
Primary Breaker	0.44	0.09	0.2	0.12	0.21	0.15	0.19	0.18	0.1	0.15	0.14	0.13
Primary Substation Bus	0.42	0.088	0.27	0.18	0.25	0.19	0.24	0.23	0.11	0.2	0.16	0.14
Primary Switch	0.27	0.23	0.3	0.25	0.6	0.51	0.54	0.52	0.43	0.38	0.39	0.34
Recloser	0.23	0.3	0.23	0.22	0.64	0.53	0.59	0.54	0.49	0.42	0.41	0.37
Primary Substation Transformer	0.096	0.26	0.11	0.13	0.35	0.31	0.35	0.3	0.34	0.2	0.26	0.2
Regulator	0.043	0.16	0.13	0.14	0.37	0.34	0.34	0.33	0.33	0.24	0.24	0.22
Elbow	0.21	0.17	0.25	0.19	0.53	0.44	0.48	0.46	0.33	0.36	0.37	0.29
Fault Indicator	0.16	0.19	0.22	0.22	0.44	0.36	0.38	0.37	0.32	0.29	0.32	0.3

Figure 4.14: Clustering of the correlation of equipment and various power outage types

Various types of outages can be grouped based on the similarity of their correlations with types of equipment in each grid cell. Based on Figure 4.14 which depicts the clustering of equipment and outage types, the outage can be categorized into three main groups. The first group with the highest correlation with the number of transformers, fuses, and conductors consists of scheduled outages, foreign interference, defective equipment, and unknown outages. The grid cells with the higher number of the equipment generally indicate the higher number of defects and requirements for maintenance, which justifies the high correlation between defective equipment and scheduled outages. The second group with a lower correlation with the number of equipments includes lightning, tree contacts, extreme wind, and wet snow. The lower correlation with the number of equipment indicates that some other factors have a more substantial influence on the number of outages, such as vegetation and weather patterns. The third group with the least correlation with the number of equipment consists of outages due to icing, adverse environment, the human element, and loss of supply. Outages due to icing seem more influenced by weather conditions rather than the number of equipment. The outages due to loss of supply have a higher correlation with the primary breaker and substation bus. Loss of supply indicates the power interruption in the transmission grid, and they are reported at the location of primary breakers, which are located at the beginning of the feeder in the substations. The adverse environment outages are mainly due to flooding and fire, which highly influence the spatial distribution of this type of outages.

4.4.2 Power Outage and Weather Data

In this section, the weather data and their relationship with power outages are investigated.

Outage probability

Power outages are dependant on various conditions such as power system protection design, topology, system state, weather condition, vegetation, environmental factors, soil type, elevation, animals, humans, etc. However, it is not feasible to measure and estimate some of the mentioned factors and conditions. In the probabilistic modeling of the power outages, we have several sources of uncertainty, including inherent stochasticity, incomplete observability, and incomplete modeling [101]. Stochasticity can appear even in deterministic systems if we cannot observe all the variables driving the behavior. For instance, we do not have the tools to measure all vegetation variations, environmental factors such as fire and flood, and the power system state due to lack of enough measurement in the distribution grid. This results in incomplete observability and adds to the uncertainty of the model. Furthermore, since we are interested in weather-related power outages, we mainly concern with finding out the interactions of weather variables with the power outages and discard some other factors that do not have any reasonable and clear interaction with weather-related power outages. For instance, we can discard the effects of humans and animals on this type of outage since we considered a separate outage cause for them.

Let us assume that system protection design, topology, equipment, and weather variables are the factors that can affect the occurrence of weather-related power outages. Bayes' theory can be used to calculate the power outage probability given that some other events such as $f_1, ..., f_n$ have happened. In (4.1), $P(outage)$ is the prior probability, $P(f_1, ..., f_n|outage)$ is the likelihood, and $P(outage|f_1, ..., f_n$ is the updated belief (posterior) of power outage with knowing extra information about $f_1, ..., f_n$.

$$P(outage|f_1, ..., f_n) = P(outage) \times \frac{P(f_1, ..., f_n|outage)}{P(f_1, ..., f_n)} \tag{4.1}$$

We can rewrite (4.1) as (4.2) and (4.3) in which the variables *weather* and *system* represent the weather condition and the power system related information such as topology, equipment, and protection.

$$P(outage|weather, system) = P(outage) \times \frac{P(weather, system|outage)}{P(weather, system)} \tag{4.2}$$

$$= P(outage) \times \frac{P(weather|outage)P(system|weather, outage)}{P(weather)P(system|weather)} \tag{4.3}$$

In (4.3) we can simplify $P(system|weather, outage)$ into $P(system|outage)$ and $P(system|weather)$ into $P(system)$ since the power system topology, number of equipment, or the protection system is independent of the weather condition. Therefore, the (4.3) will be simplified into (4.4).

$$P(outage|weather, system) = P(outage) \times \frac{P(weather|outage)P(system|outage)}{P(weather)P(system)} \tag{4.4}$$

We can take advantage of available spatial information and define a unique probabilistic model for each grid cell in the service area. The location information embed some information about its unique characteristics such as vegetation, environmental factors, elevation, and power system related information. However, by spatial discretization of the service area, we discard some information about the precise location of events and equipment, which will add some uncertainty due to incomplete modeling. We consider the location by defining a unique probability distribution for outages for each grid cell. Therefore, variables such as the number of equipment and system topology can be considered independent of power outages since they are the same during power outages or after power restoration. Therefore $P(system|outage)$ can be written as $P(system)$, and $P(outage|weather, system)$ as $P(outage|weather)$ for each grid cell and (4.4) can be written as:

$$P(outage_{(x,y)}|weather_{(x,y)}) = P(outage_{(x,y)}) \times \frac{P(weather_{(x,y)}|outage_{(x,y)})}{P(weather_{(x,y)})} \tag{4.5}$$

The $\mathrm{o}_{(x,y)}$ is a random variable that can take different values randomly which represents the number of weather-related power outages during a time window in a specific grid cell. The sample space for the outcome of $\mathrm{o}_{(x,y)}$ can be defined as $\Omega = \{0, 1, 2, 3, ...\}$ that represents the set of number of power outages. The events $o_0 = \{0\}$, and $o_1 = \{1, 2, 3, ...\}$ are two members of the event space (ξ) which represent having no outage or having at least one outage in a time window. The event o_1 is the complement of o_0 and to satisfy the axiom of probability $P(\Omega) = 1$, we get $P(o_1) + P(o_0) = 1$. The $\mathrm{W}_{(x,y)} = (W_1, W_2, ..., W_d)_{(x,y)}$ is a multivariate random variable which is a vector of random variables with vector valued outcomes $\mathbf{w} = (w_1, w_2, ..., w_d)$ representing weather condition features. Equation (4.6) estimates the probability of having at least one weather-related power outage in a $\text{gridcell}_{(x,y)}$ over a specific time window and having weather condition in the range of $\mathbf{w_1}$ and $\mathbf{w_2}$.

$$\begin{aligned} &P(\mathrm{o}_{(x,y)} = o_1|\mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2}) = \\ &\qquad P(\mathrm{o}_{(x,y)} = o_1) \times \frac{P(\mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2}|\mathrm{o}_{(x,y)} = o_1)}{P(\mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2})} \end{aligned} \tag{4.6}$$

In order to calculate the power outage probability, we use a sliding window with length T to create samples and slide it with stride S over four years of weather and power outage data. Information in each window is considered as one sample from the sample space. The prior probability of power outage for gridcell$_{(x,y)}$ can be calculated as:

$$P(\mathrm{o}_{(x,y)} = o_1) = \frac{\text{num windows with outage in cell}_{(x,y)}}{\text{num windows in the sample space in cell}_{(x,y)}} \tag{4.7}$$

calculate the likelihood of weather conditions given power outage as:

$$P(\mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2} | \mathrm{o}_{(x,y)} = o_1) = \frac{\text{num windows with outage in cell}_{(x,y)} \text{ that satisfies the weather condition}}{\text{num windows with outage in cell}_{(x,y)}} \tag{4.8}$$

and calculate the normalization factor as:

$$P(\mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2}) = \frac{\text{num windows in cell}_{(x,y)} \text{ that satisfies the weather condition}}{\text{num windows in the sample space in cell}_{(x,y)}} \tag{4.9}$$

By substituting (4.7), (4.8), and (4.9) into (4.6) we get:

$$P(\mathrm{o}_{(x,y)} = o_1 | \mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2}) = \frac{\text{num windows with outage in cell}_{(x,y)} \text{ that satisfies the weather condition}}{\text{num windows in cell}_{(x,y)} \text{ that satisfies the weather condition}} \tag{4.10}$$

The (4.10) can be calculated as:

$$P(\mathrm{o}_{(x,y)} = o_1 | \mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2}) = \frac{\sum_{n=1}^{N_{window}} o_1\_condition_n^{(x,y)}}{\sum_{n=1}^{N_{window}} condition_n^{(x,y)}} \tag{4.11}$$

where

$$condition_n^{(x,y)} = \sum_{i=1}^{T} \omega_i F_{condition}^{(x,y)}(t_{n_i}) \tag{4.12}$$

$$o_1\_condition_n^{(x,y)} = \sum_{i=1}^{T} \omega_i F_{condition}^{(x,y)}(t_{n_i}) F_{o_1}^{(x,y)}(t_{n_i}) \tag{4.13}$$

The functions $F_{condition}^{(x,y)}$, and $F_{o_1}^{(x,y)}$ are defined to determine how much each window satisfies the conditions. The $F_{condition}^{(x,y)}$ as defined in (4.14) returns 1 for any hour in the time window that the weather condition in $gridcell_{(x,y)}$ is satisfied, otherwise it returns zero. The $F_{o_1A}^{(x,y)}$ as defined in (4.15) returns one for every hour before the last power outage in a given time window in $gridcell_{(x,y)}$, and it returns zero after the last power outage in the time window. The $F_{o_1A}^{(x,y)}$ is defined based on the fact that only the weather condition before the outage contribute to the power outage event. However, one may argue that $F_{o_1A}^{(x,y)}$ can underestimate the power outage probability, and the weather condition after the power outage should also be considered. When power outages happen, the faulty part of the power system gets isolated; therefore, even if the weather condition is the same as before the power outage and potentially can cause more power outages, the utility may not experience any other power outages downstream of the grid. Therefore, we introduce $F_{o_1B}^{(x,y)}$ as defined in (4.16) which returns one for every hour in a time window if a power outage occur in that window. However, it should be considered that $F_{o_1B}^{(x,y)}$ may overestimate the power outage probability.

$$F_{condition}^{(x,y)}(t_{n_i}) = \begin{cases} 1 & \text{if} \quad \mathbf{w_1} < \mathrm{W}_{(x,y)t_{n_i}} < \mathbf{w_2} \\ 0 & \text{if} \quad \text{else} \end{cases} \tag{4.14}$$

$$F_{o_1A}^{(x,y)}(t_{n_i}) = \begin{cases} 1 & \text{if} \quad t_{n_1} \le t_{n_i} \le max(t_{(x,y)outage}) \le t_{n_1} + T \\ 0 & \text{if} \quad t_{n_1} \le max(t_{(x,y)outage}) \le t_i \le t_{n_1} + T \end{cases} \tag{4.15}$$

$$F_{o_1B}^{(x,y)}(t_{n_i}) = \begin{cases} 1 & \text{if} \quad t_{n_1} \le t_{(x,y)outage} \le t_{n_1} + T \\ 0 & \text{if} \quad \text{else} \end{cases} \tag{4.16}$$

Having storms or specific weather conditions just before the power outage may seem more important than satisfying the same conditions several hours after the power outages. Therefore, we defined the weights ω_i corresponding to the hours in the window to assign different weights for the each hour. In order to limit the value of $condition_n^{(x,y)}$ and $o_1\_condition_n^{(x,y)}$, the constraint (4.17) is defined to limit the maximum value of each window to one.

$$\sum_{i=1}^{T} \omega_i = 1 \tag{4.17}$$

By defining $F_{o_0A}^{(x,y)}(t_{n_i})$ as (4.18), and $F_{o_0B}^{(x,y)}(t_{n_i})$ as (4.19) and substituting each of them into (4.13) we can verify that (4.20) and (4.21) are satisfied.

$$F_{o_0A}^{(x,y)}(t_{n_i}) = \begin{cases} 0 & \text{if} \quad t_{n_1} \leq t_{n_i} \leq max(t_{(x,y)outage}) \leq t_{n_1} + T \\ 1 & \text{if} \quad t_{n_1} \leq max(t_{(x,y)outage}) \leq t_i \leq t_{n_1} + T \end{cases} \tag{4.18}$$

$$F_{o_0B}^{(x,y)}(t_{n_i}) = \begin{cases} 0 & \text{if} \quad t_{n_1} \leq t_{(x,y)outage} \leq t_{n_1} + T \\ 1 & \text{if} \quad \text{else} \end{cases} \tag{4.19}$$

$$o_1\_condition_n^{(x,y)} + o_0\_condition_n^{(x,y)} = condition_n^{(x,y)} \tag{4.20}$$

$$P(\mathrm{o}_{(x,y)} = o_1 | \mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2}) + P(\mathrm{o}_{(x,y)} = o_0 | \mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2}) = 1 \tag{4.21}$$

In Figure 4.15 the weather-related power outage probabilities across one service area with weather conditions defined in (4.22), and parameters defined in (4.23) is depicted. For each grid cell, the power outage probabilities are calculated using (4.11).

$$\begin{gathered} \text{wind speed} > 10\, km/h, \\ \text{temperature} < -5, \end{gathered} \tag{4.22}$$

$$\begin{gathered} T = 3,\ S = 1, \\ [\omega_1, \omega_2, \omega_3] = [0.1, 0.3, 0.6] \end{gathered} \tag{4.23}$$

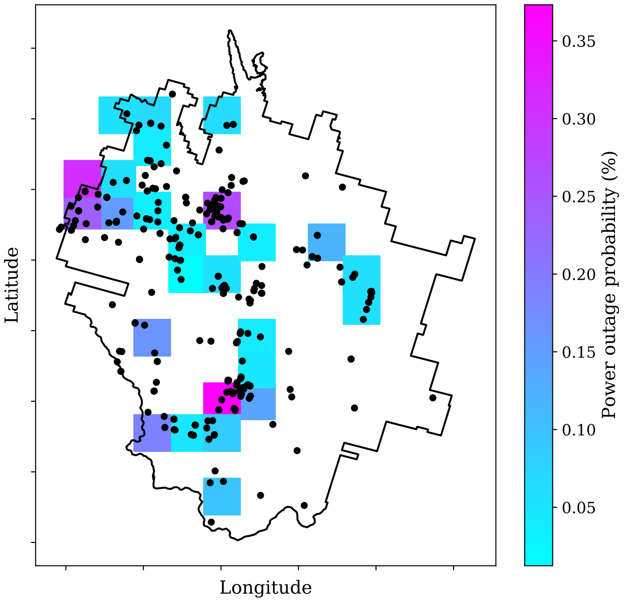


Figure 4.15: Power outage probability given specific weather condition

Power utilities mainly depend on the skills of their experienced crews to estimate the chance of having power outages based on their knowledge about the past storms and consequent outages happened in the grid. However, the proposed method enables the power utility to utilize their massive amount of power outage data in their OMS system. Based on the historical events, they can estimate the probability of power outage on a grid cell level and update that estimation with the newly available weather forecast data. Furthermore, this method can be utilized to investigate other scenarios like having 10% stronger or weaker storms and update the estimation based on past events. It can also be used to identify the more sensitive locations to specific weather conditions and reveal the locations that historically were more prone to power outages. However, it should be noted that this method relies on the historical power outages that happened in one area, and it will not provide any valuable probability estimation if there is not any power outage in the database that can satisfy the defined conditions. Although it can be considered a weakness that this method cannot extrapolate and interpolate to provide estimations when the data is not available, with the recent advances in the data-driven methods, more utilities are continuously collecting data that can highly improve the future probability estimation.

Classification

In this section, we define a classification model to predict power outages given the weather condition. The goal is to minimize the expected cost of the classifier. The cost$(\hat{o}, o)$ represents the penalty or cost of predicting $\hat{o}$ where the true target is o. Given continuous weather vectors as inputs from ω, and discrete target set $\{o_0, o_1\}$ we can express the expected cost as follows:

$$E[cost] = \int_{\omega} \sum_{o \in \{o_0,o_1\}} cost(\hat{o}, o) p(\mathbf{w}, o) d\mathbf{w} \tag{4.24}$$

$$= \int_{\omega} p(\mathbf{w}) \sum_{o \in \{o_0,o_1\}} cost(\hat{o}, o) p(o|\mathbf{w}) d\mathbf{w} \tag{4.25}$$

In optimal classifier we want to minimize the expected cost; therefore, the optimal classifier can be written as:

$$\underset{\hat{o} \in \{o_0,o_1\}}{\text{argmin}} \; E(cost|\text{W} = \mathbf{w}) \tag{4.26}$$

For each input $\mathbf{w}$ the optimal classifier can be expressed as (4.27):

$$f^*(\mathbf{w}) = \underset{\hat{o} \in \{o_0,o_1\}}{\text{argmin}} \sum_{o \in \{o_0,o_1\}} cost(\hat{o}, o) p(o|\mathbf{w}) \tag{4.27}$$

For a cost function defined as:

$$cost(\hat{o}, o) = \begin{cases} 0 & \text{if} \quad o = \hat{o} \\ 1 & \text{if} \quad o \neq \hat{o} \end{cases} \tag{4.28}$$

By converting argmin into argmax and knowing $\sum_{o \in \{o_0,o_1\}} p(o|\mathbf{w}) = 1$ the f^* can be simplified into:

$$f^*(\mathbf{w}) = \underset{\hat{o} \in \{o_0,o_1\}}{\text{argmax}} \, (1 - \sum_{o \in \{o_0,o_1\}} cost(\hat{o}, o) p(o|\mathbf{w})) \tag{4.29}$$

$$= \underset{\hat{o} \in \{o_0,o_1\}}{\text{argmax}} \sum_{o \in \{o_0,o_1\}} (1 - cost(\hat{o}, o)) p(o|\mathbf{w}) \tag{4.30}$$

$$= \underset{\hat{o} \in \{o_0,o_1\}}{\text{argmax}} \sum_{o \in \{o_0,o_1\}, o \neq \hat{o}} (1 - 1).p(o|\mathbf{w}) + \sum_{o \in \{o_0,o_1\}, o = \hat{o}} (1 - 0).p(o|\mathbf{w}) \tag{4.31}$$

$$= \underset{o \in \{o_0,o_1\}}{\text{argmax}} \, p(o|\mathbf{w}) \tag{4.32}$$

Therefore, by calculating posterior distribution $P(\mathrm{o}_{(x,y)} = o_0|\mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2})$ and $P(\mathrm{o}_{(x,y)} = o_1|\mathbf{w_1} < \mathrm{W}_{(x,y)} < \mathbf{w_2})$ and choosing the highest probability we can minimize the expected cost. However, it should be noted that the power outages are rare events compared to the no outage conditions. Therefore, the definition of cost function should be modified to compensate the rareness and higher importance of power outages.

In general, however, the probability distribution $p(o|\mathbf{w})$ is usually approximated using a specific functional form and a set of adjustable parameters. The direct estimation of posterior probability could not be possible if there is not enough historical data. However, the functional estimation of the probability distribution enables us to estimate the posterior probability even for the inputs that do not correspond to any data point in the historical power outages and weather conditions.

4.5 Conclusion

This chapter provides a detailed overview of the outage management systems that utilities use for storing and analysis of outage data. We investigate and explain the relationships between various data sources, and demonstrate results of integration of the power outage, power system, and weather data. New insights and statistics regarding the power outage causes, protective devices, duration, time, and location are presented. The relationship between power equipment and power outage types is investigated, and based on their similarity, three main groups of power outages are identified.

We utilize Bayes' rule and define posterior probability distribution for power outages. We introduce a window sampling for the power outage probability calculation. The proposed analysis method can be used by power utilities to perform their assessment of power outages based on the available weather data. Furthermore, the maximum posterior classifier model for power outage prediction is derived. We demonstrate that it can be used as a classifier to distinguish power outages versus no power outage conditions. At the same time, we show that it may not be applicable to use it as a classifier when there are not enough data points.

Chapter 5

Weather Outage Prediction system – ***WoutPS***

Improvements in monitoring and data collection practices provide opportunities for more comprehensive modelling and managing grid operations. At the same time, advanced data analysis methods should be able to address service quality degradation due to outages, weather patterns and asset-related performance. In this chapter, we apply Machine Learning and Computational Intelligence methods for the analysis of power distribution system data and constructing a system for predicting power outages. Weather and outage data are utilized by the proposed system for predicting purposes. We evaluate the prediction performance of different types of prediction models. We also propose and validate three different architectures of a system for predicting types of weather-related outages. We focus on outages caused by wind, snow and ice. An analysis of the prediction results is provided.

5.1 Introduction

Power outages in distribution networks are relatively frequent, and they impose a high cost on power utilities and considerable inconvenience to customers. According to [1], 44%-78% of the power outages reported in various studies were weather-related, costing between $20 billion and $55 billion annually only to the U.S. economy. A utility usually needs to send out a large number of crews to restore services during power outages, followed by severe weather conditions. The estimated cost of an average storm is around $100,000 to $1,000,000 per hour [2].

Prediction of weather-related power outages and their severity utilization have two aspects. First, it can be utilized to identify the most vulnerable locations to extreme weather in the power grid and to suggest long term resilience investment and hardening programs. These long term actions will influence grid resistance, reliability, and redundancy. Second, it will be utilized for rapid recovery after outages, and contingency plans and emergency operations [10]. Rapid recovery plans play a significant role in preventing very costly outages by helping utilities to forward plan the resource allocation prior to the outage. Outage prediction results in increased profitability of utilities, improved grid reliability, resiliency, operational efficiency, and enhanced customer experience [11].

In this chapter, a collection of data-driven algorithms and methods that constitute a framework suitable for building a system for predicting weather-related outages is developed. The core methodology used to build the framework consist of technologies, including Machine Learning algorithms and elements of Computational Intelligence, which results in:

- developing models suitable for predicting different types of weather-related outages;
- developing a novel approach for reasoning under uncertainty to combine evidence from various models based on Dempster-Shafer theory (DST);
- constructing a *Weather outage Prediction System* (*WoutPS*) based on multiple models combined with reasoning framework for predicting outages.

A significant number of research papers related to outage analysis focus on models for predicting power outages caused by storms such as hurricanes, blizzards, tornadoes, and thunderstorms. However, outages due to wet snow and icing have not gained a lot of attention in the literature. The proposed models and the system *WoutPS* are able to distinguish three main causes of outages, including severe wind, wet snow, and icing. Furthermore, in order to provide users with more information so decisions made using *WoutPS* are better understood, a novel DST-based aggregation method is proposed to provide levels of confidence in *WoutPS*'s predictions. This method combines the results of individual classifiers and leads to a more powerful ensemble model.

5.2 Related Work

Outages, as disruptions in energy delivery, are quite frequent events [102], [103]. They happen more often in systems with ageing equipment, as well as due to changing weather patterns [104]. The ability to predict their occurrence and severity is essential for power utilities [105]. Multiple different methods and models, that include SVM [106], linear regression [107], artificial neural networks [108], and tree-based models [26], [30], have been reported in literature. Most of them are based on analysis of historical data about outages, weather and environment [9].

Parametric statistical models such as negative binomial generalized linear model and Poisson generalized additive model are developed for estimation of the spatial distribution of power outages caused by hurricane [109], [110]. Non-parametric models such as Bayesian additive regression splines and classification and regression trees further improved the prediction accuracy compared to non-parametric models [111].

The prediction of number of outages due to extreme weather events – ice storms, hurricanes – has been addressed in [2], [112]–[114]. An interesting work focusing on predicting wind and lightning outages using artificial neural networks with a modified back-propagation algorithm has been presented in [115]. Various extreme weather events such as thunderstorms, hurricanes, and blizzards are considered to develop prediction models in [30]. In [112], two models for power outage estimation are developed to predict the overhead lines failure rate in the distribution grid caused by ice/snow storms and thunderstorms. The failure rate has been predicted based on average wind speed and different weather categories using models such as multiple linear regression [112] and generalized linear models [116]. A Poisson regression model combined with a Bayesian network model using gust wind speed and lightning strike counts have also been reported in [112]. A fuzzy clustering-based approach has been used to predict the impact of hurricanes on the failure rate of transmission lines [117]. The model has been built on data representing wind speed and rainfall. The outage prediction model built based on gust wind speed, duration of strong winds, week-long rainfall, and population density has been described in [118].

Another work of line developed models to rank power system components by their susceptibility to failure [119], [120]. Ensemble models based on boosting algorithms outperformed the accuracy of neural networks [115] and a mixture of expert models [121] for prediction of lighting and wind-related power outages [122]. One of the most effective ensemble models used for outage prediction has been random forest (RF), which is able to capture the non-linearity of complex outage data. It has been used in [123] and utilized input data such as a wind gust speed, a duration of wind above 20 m/s, a number of customers, tree trimming, and soil moisture. Similarly, in [32], the authors have presented an RF predictor built based on wind, soil and drought data, as well as eight classes of land cover to improve prediction performance. In [29], the same type of models have been used to predict outages due to hurricanes; vegetation and location have been used as input data.

Hybrid classification method, tree/regression, and its ability to handle zero-inflation is investigated in [124]; furthermore, quantile regression forests coupled with RF model is developed to give more insight about target variable distribution [125], and to demonstrate the usefulness of land cover variables in predicting power outages [126]. It further used Monte Carlo simulation to investigate the tropical cyclone track impact on prediction performance, and demonstrated the importance of initial intensity and official track forecast.

Other models – based on fuzzy logic techniques and neural networks – have been presented in [127] and [128], respectively. The fuzzy model has evaluated weather-related hazards and probabilities of occurrence of component outages, while the neural network one has predicted outage duration combining environmental factors with textual information from field reports.

5.3 Data Description

The final performance of any arbitrarily optimal model, even with an unlimited number of data, cannot exceed a certain value [129]. The reason is an inherent, irreducible error, which is the minimum error that can be achieved. Irreducible error is the result of lack of information, partial observability, noise, and variability of the target variable, leading to a distribution on the probable target variable for any input data [130].

A process of constructing machine learning models requires access to a sufficient amount of high-quality data. Data is an integral element of building outage prediction models. In this research, a great effort is put into collecting and pre-processing the required data.

5.3.1 Outage Data Description

At most power utilities, outage data are stored in the form of relational databases called Outage Management Systems (OMSs) (Section 4.2). The OMS records included an outage start time, a restoration time, geographic coordinates, number of affected customers, affected feeder, transformers, and an isolating device. An outage start time is considered the time when the first customer reports the outage. An outage location is approximated, the location of the first upstream isolating device, i.e., overcurrent relays, switches, reclosers, fuses, and transformers. When several customers report an outage, the outage location is updated considering the location of each customer. Nested outages are distinguished as they have the same start time but different restoration times and they are considered as unique outages.

Outage types are categorized into nine groups: unknown, scheduled outage, tree contacts, lightning, defective equipment, extreme weather, adverse environment, the human element, and foreign interference. In this research, we focus on predicting outages caused by extreme weather conditions.

5.3.2 Weather Data Description

Extreme weather conditions identified here refer to: extreme wind, wet snow accumulation, and equipment icing caused by wind/winter storms. The effect of extreme wind on a power system condition is substantial. The extreme wind breaks the trees and causes them to collapse on the power lines. Throughout the spring and summer times, when leaves are on trees even not so, strong winds lead to power outages by breaking the branches and throwing them over power lines [131]. Extreme winds can also lead to the collapse of poles depending on a type of pole, a level of maintenance, a span length, and deterioration caused by age or insects [132], [133].

Both ice accretion and snow accumulated over power equipment cause several problems. Some equipment fails due to an icing flashover [134]–[136], and when combined with extreme wind icing can cause a phenomena called 'galloping' in overhead conductors [137], [138]. Ice accretion on equipment is influenced by elevation and atmospheric conditions such as air temperature and wind speed. Its growth rate is dependant on a water droplet velocity, size, and effective structure area [139], [140]. A flashover is more likely to occur at higher voltages due to decreased dielectric strength of ice-covered insulators [141]. Decreased impedance leads to increased voltage stress over air-gaps, and creates partial arcs which can grow into high energy arcs leading to a flashover [134], [135], [142], [143]. Ice and freezing rain also accumulate on power lines and form heavy layers of ice around conductors. This leads to collapsed power lines. Ice also can accumulate on tree limbs. The increased weight damages trees and causes them to collapse on power lines [131]. Wet snow, similarly to icing and freezing rain, causes trees to collapse on power lines. Moreover, it can lead to the breakdown of insulation components.

Wind-related parameters are used because a wind-force is the primary reason for tree uprooting or breakage, as well as the collapse of poles. Precipitation moistens the ground and leads to uprooting trees uprooting in the presence of high winds [28]. Furthermore, precipitation in the form of snow, as well as ice, temperature and humidity, are the main atmospheric factors leading to icing and wet snow outages.

The weather parameters used in this study include air temperature (ºC), relative humidity (%), wind speed (km/hr), wind direction (º), surface pressure (kPa), mean sea level pressure (kPa), cloud coverage (%), and precipitation (mm).

5.3.3 Importance of Parameters for Analysis of Outage

The important weather parameters affecting power outages are investigated in this section based on univariate statistical testing. Statistical hypothesis tests, such as ANOVA (Analysis of Variance), are widely used in analyzing and understanding the difference between group means. ANOVA test is primarily used to determine whether the means of two or more populations are equal or there are statistically significant differences between them. The null hypothesis is that the group means are equal. For instance, we want to see if the mean value of weather parameters are the same for various types of power outages. If we can reject the null hypothesis, we can conclude that the average value of weather parameters significantly varies between different power outage types and could be an important feature for classifying power outages and their types.

Table 5.1 presents the F-statistic and P-values for various weather parameters from four groups of data, including power outages caused by extreme wind, wet snow, icing, and no outages. It can be observed that the mean value of weather parameters for various types of power outages are not the same since the P-values are less than 0.05 level of significance and the null hypothesis can be rejected with a confidence interval of 95%. In Table 5.2, based on the results of the ANOVA test for each type of power outage versus no outage, the weather parameters are listed in order of their importance.

5.4 Weather Outage Prediction System – ***WoutPS***

A model that can predict the expected value of the target variable given an input data point is an optimal model and has the highest performance given

Table 5.1: F-statistic and P-value for weather parameters

Feature	F-statistic	P-value
Temperature	1557.44	0.0
Snowfall	1463.03	0.0
Relative humidity	893.99	0.0
Wind speed	514.41	0.0
Cloud coverage	293.47	1.40 e-187
Surface pressure	249.48	8.98 e-160
Precipitation	162.98	9.79 e-105
Mean sea level pressure	118.02	5.97 e-76
Wind direction	112.75	1.42 e-72

Table 5.2: Weather parameters importance on the power outage types

Importance	Extreme wind	Wet snow	Icing
1	Wind speed	Snowfall	Temperature
2	Surface pressure	Relative humidity	Relative humidity
3	Wind direction	Temperature	Surface pressure
4	Mean sea level pressure	Cloud coverage	Mean sea level pressure
5	Precipitation	Precipitation	Precipitation
6	Temperature	Wind direction	Cloud coverage
7	Cloud coverage	Mean sea level pressure	Wind direction
8	Snowfall	Surface pressure	Wind speed
9	Relative humidity	Wind speed	Snowfall

a data-set. In order to build an optimal model, two components of a prediction error, i.e., bias and variance are considered. Simple models that are not powerful enough to represent a true input-output relation introduce some bias and low variance, which is called under-fitting. On the other hand, too complex models usually correspond to a true input-output relation, but they lead to over-fitting to the training data. Complex models can fit the noise, and consequently, they may not generalize well. Both of them are a function of model complexity and can be minimized by a proper selection of model hyper-parameters. A random search approach is implemented to find the best combination of hyperparameters, i.e. a single model is built and evaluated for each combination of model hyper-parameters.

5.4.1 System Architecture Investigations

In order to ensure a balance between the simplicity and complexity of a system yet to obtain good performance results, we investigate three architectures for predicting extreme weather power outages and their causes.

Architecture *a*: is a simple multi-class classifier based on weather parameters and location information recognizes and predicts the following conditions: NoO (NoOutage), O (outage) – extreme wind outage, wet snow outage, and icing caused outages. Classifiers such as RF, MLP, KN, DT are inherently capable of multi-class classification. However, classifiers such as SVC require a one-vs-all strategy to reduce a multi-class problem to multiple binary prediction problems.

Architecture *b*: splits the prediction process into two stages. At the first stage, we have a binary classification that distinguishes NoO versus O conditions. At the second stage, we have a multi-class predictor that is able to identify the outage cause – extreme wind, wet snow, and icing.

Architecture *c*: is kind of a combination of the two previously presented architectures. There are two motivations here. Firstly, an optimal model may not be constructed using a specific classification algorithm and its tuneable hyper-parameters. It can be proved that an ensemble style optimal classifier has the best performance. Both these facts lead us to construct various prediction models and combining their outputs. Therefore, the architecture c leverages the power of ensemble learning. Secondly, each classifier is considered as a piece of evidence, and DST is utilized to determine a confidence level in the final prediction. The level of confidence in each classifier is determined by calculating a mean of cross-validated F_1 scores of the classifier with best hyper-parameters, Algorithm 1.

5.4.2 Data Pre-processing and Model Construction

Collected weather data are standardized, i.e., the distribution values are rescaled to have the mean values of zero and standard deviation of one. Stan-

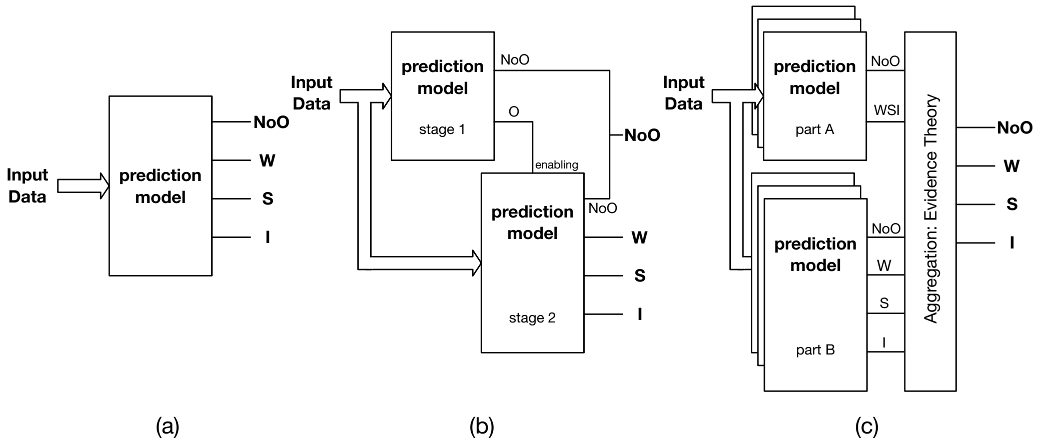


Figure 5.1: System Architecture [144] ©2020 IEEE

Algorithm 1 Fusion algorithm, [144] ©2020 IEEE

```
Initialize classifiers $Ci, j : i \in \{RF, MLP, SVC\}, j \in \{A_n, B_n\}$
Initialize frame of discernment $\Theta = \{wind, snow, ice, NoO\}$ and its power set $2^{\Theta}$
$A_n : n \in \{\{NoO\}, \{wind, snow, ice\}\}$
$B_n : n \in \{\{NoO\}, \{wind\}, \{snow\}, \{ice\}\}$
for $i, j$ do
    $c_{i,j} \leftarrow C_{i,j}$ with best mean cross-validated $F_1$ score
    $confidence_{c_{i,j}} \leftarrow mean(F_1(c_{i,j}))$
procedure FUSION($c_{i,j}, confidence_{c_{i,j}}$)
    for $i, j$ do
        $m_{c_{i,j},ignorance} \leftarrow 1 - confidence_{c_{i,j}}$
        for $n$ do
            $m_{c_{i,j},n} \leftarrow Pr_{c_{i,j},n} \times confidence_{c_{i,j}}$
    $m = \bigoplus_{i,j} m_{i,j}$
    $Pr_a = BetPm(a), \quad \forall a \in \Theta$
    return $Pr_a, \quad \forall a \in \Theta$
```

dardization is a feature scaling method that plays an essential role in training prediction models by helping gradient descent to converge faster. Z-score, which is a parametric outlier detection method that assumes a Gaussian distribution for weather parameters, is implemented to detect and remove samples that are 'far' from sample mean values. These samples could be a result of a wrong measurement process, make data unreliable. This profoundly affects the model performance.

The models used in our *WoutPS* predict the occurrence and type of an outage based on 12 hours of weather data prior to the outage time. These models extract location-sensitive patterns in weather data that lead to an outage. Assuming weather parameters in the past are available at each time, power outages can be predicted up to 12 hours prior to its occurrence. For instance, to predict a probability of power outage 8 hours into the future, 4 hours of weather data in the past and 8 hours of weather forecast are required.

In order to construct the best possible model, a number of evaluation techniques are used. They are also applied to determine the values of hyperparameters of a single model and to compare different models. To properly compare built prediction models, we should ensure that performance metrics obtained for the available data are reliable indicators of the model performance when applied to new data. To achieve this confidence, stratified k-fold cross-validation (CV) and statistical significance tests are implemented to gain evidence in the model's prediction abilities. K-fold CV splits the data into k subsets, and a single model is trained on k-1 folds, named training partition, then evaluated on the holdout validation partition. The processes of partitioning the data, training the model and evaluating it are repeated for k times on k different validation partitions and their associated training partitions, Fig. 5.2. The cross-validation technique is less likely to add bias to the model since all samples have the chance to participate in training and evaluation processes. Stratified k-fold CV is a technique that ensures each fold has a similar distribution to the data by keeping the same percentage of each class across each fold by rearranging the data. When the data is unbalanced, stratification plays a significant role in preventing the over-representation of a specific class. Implementation of CV provides k performance measures obtained for each model.

5.5 Outage Prediction Models

The measure used to evaluate the models' performance is F_1. So, k F_1 are used to find out whether a difference in the performance of two models is meaningful or is due to a chance. The models' difference can be proved by rejecting the null hypothesis H_0, which assumes two models are equal. The null hypothesis

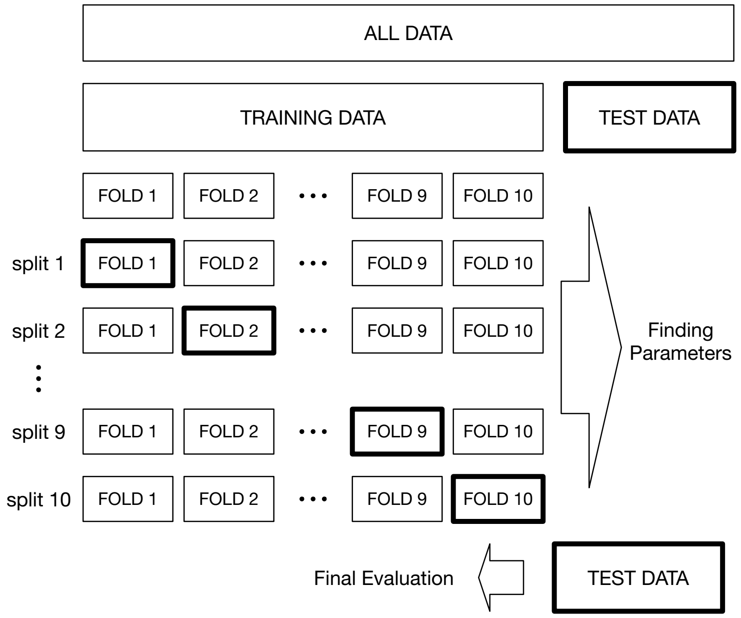


Figure 5.2: Cross validation process [145], [144] ©2020 IEEE

can be rejected if $p \leq \alpha$, i.e., p value is the probability $Pr(T > t = \frac{\bar{d}-0}{\sigma_d/\sqrt{n}})$, where T is a random variable, σ_d and $\bar{d}$ are standard deviation and mean of F_1 measures of each model, and α is the significance value which is usually assumed to be equal to 0.05. If H_0 cannot be rejected, it can be concluded that there is not sufficient evidence that the two models are different.

5.5.1 Model Selection Process

Various outage prediction models have been evaluated to select a set of models that provided the best prediction results. We have used a random search approach with various combinations of hyper-parameters for each model, and for each combination, a stratified 10-fold CV has been implemented on 80% of the data, which is sampled randomly for training and testing purposes. Best hyper-parameters are selected for each model by comparing all combinations, Fig. 5.3. The obtained results for all models used in binary and multi-class classifiers with tuned hyper-parameters are depicted in Fig. 5.4 and 5.5, respectively.

RF, MLP, and SVC classifiers have the best F_1 scores among all evaluated classifiers. However, applying independent t-test on CV results demonstrated

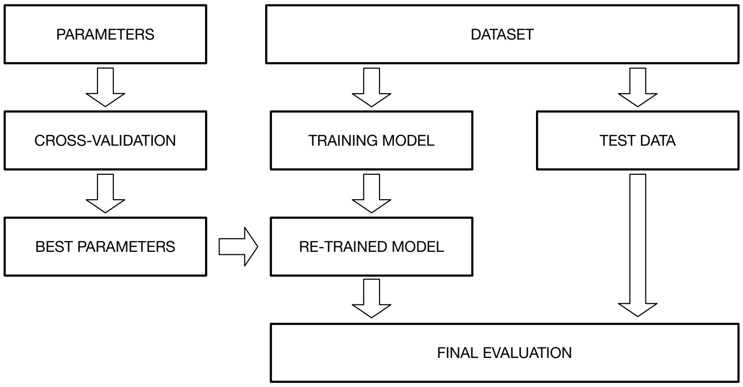


Figure 5.3: Hyper-parameters tuning [145], [144] ©2020 IEEE

that there is no sufficient evidence that the performances of these classifiers are statistically different, Table. 5.3. Therefore, a system with the architecture a is trained with only an SVC classifier, which has the highest F_1 score.

Table 5.3: P-values for student T-test for all model, [144] ©2020 IEEE

Model	RF	MLP	SVC	KN	AdBo	DT	QDA
RF	1	0.7915	0.2939	0.0001	3.4941 e^{-8}	2.3690 e^{-7}	1.2680 e^{-8}
MLP		1	0.1898	0.0001	2.9700 e^{-8}	1.8938 e^{-7}	1.0970 e^{-8}
SVC			1	0.0004	1.5844e^{-8}	1.9914 e^{-7}	1.6626 e^{-8}
KN				1	0.0212	0.04532	3.3652 e^{-6}
AdBo					1	0.7963	2.4009 e^{-5}
DT						1	2.4830 e^{-5}
QDA							1

5.5.2 Evaluation of *WoutPS*

Based on the results of identifying the best prediction model, we conclude that *WoutPS* can be built with any of the three best models: RF, MLP, and SVC. Therefore, it seems that the selection of a specific classier among these three should not affect the process of selecting the most suitable architecture.

An evaluation process is done using 20% of the available data. Please, recall we have used 80% of data for CV process of evaluating the best prediction model. The results of the evaluation are presented in Table 5.4. The values in bold represent the best precision, recall, and F_1 obtained across all three

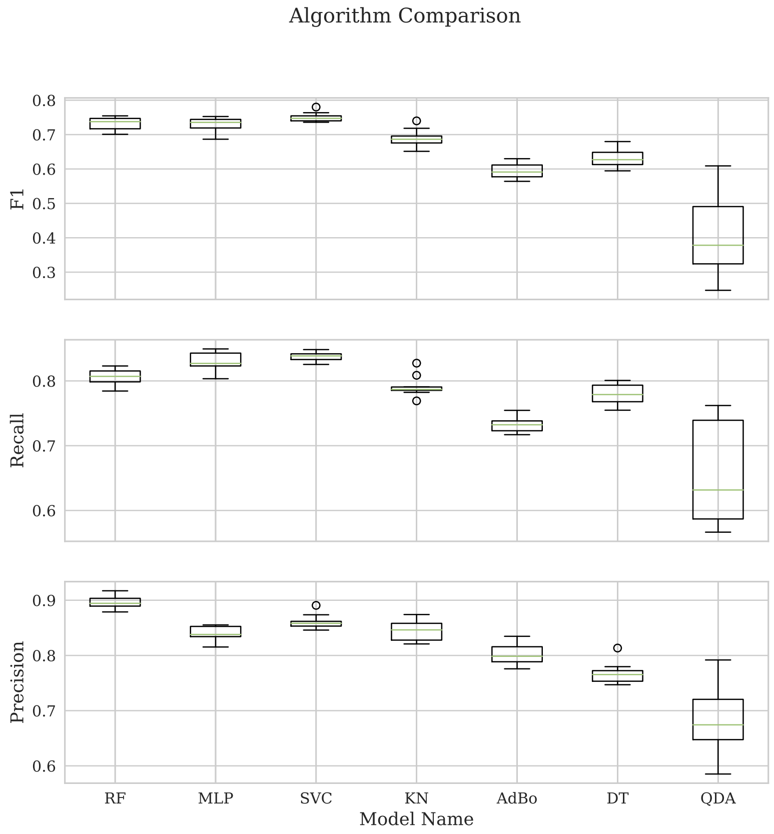


Figure 5.4: Comparison of macro average $F_1, recall, precision$ score of binary classifier (stage 1) used in architecture b, c

architectures. As can be observed, *WoutPS* with the architecture c has the majority of bold, i.e., best scores. Only the value of recall for predicting Icing is better for architecture a, while the value of precision for NoO is the same for all architectures. Such a result is a clear indication of better performance when the architecture c is used.

Furthermore, if we look into the precision and recall performance metrics in 5.4 and 5.5, it can be observed that the SVC has the highest recall, and the RF has the highest precision among the models. It seems architecture c which utilizes the DST-based aggregation of RF, SVC, and MLP models, can take advantage of each unique model's better recall and precision and provide a more powerful ensemble model.

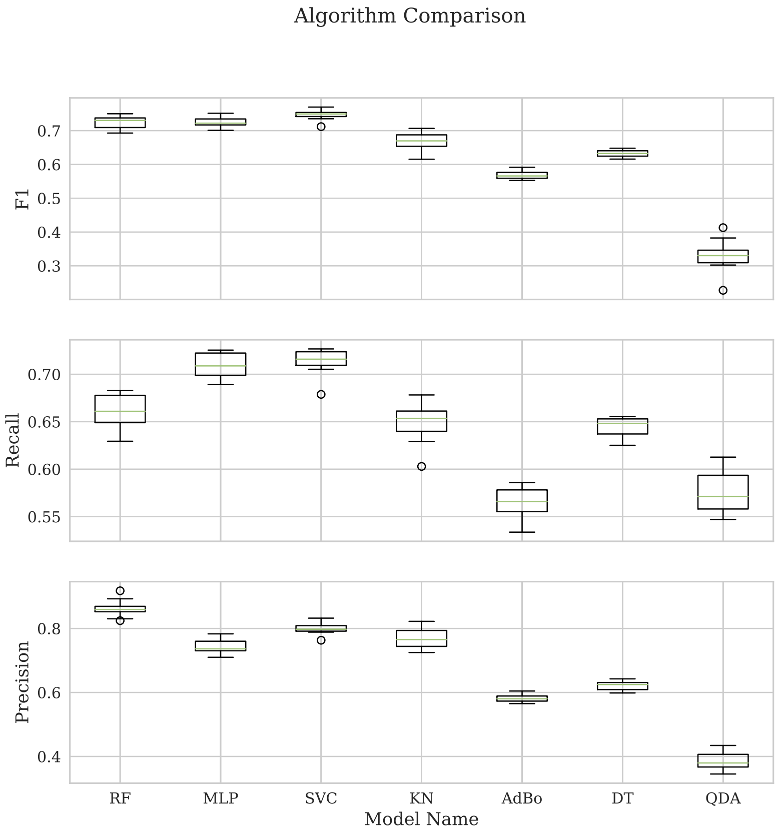


Figure 5.5: Comparison of macro average $F_1, recall, precision$ score of multi-class (stage 2) classifier used in all architectures.

Table 5.4: Precision, Recall, and F_1 score report on test data for architectures a, b, c, [144] ©2020 IEEE

Metric	NoO	Wind	Snow	Icing	Macro Ave.
Precision a	0.93	0.65	0.83	0.73	0.79
Recall a	0.96	0.41	0.68	**0.76**	0.70
F_1 score a	0.94	0.50	0.75	0.74	0.74
Precision b	0.93	0.66	0.85	0.74	0.80
Recall b	0.96	0.41	0.68	0.73	0.70
F_1 score b	0.95	0.51	0.76	0.74	0.74
Precision c	0.93	**0.68**	**0.86**	**0.81**	**0.82**
Recall c	**0.97**	**0.42**	**0.73**	0.73	**0.71**
F_1 score c	**0.95**	**0.52**	**0.79**	**0.77**	**0.76**

5.5.3 *WoutPS* with Dempster-Shafer Aggregation

As it has been indicated earlier, the best results are obtained for *WoutPS* with the architecture *c*. The application of the Demeter-Shafer theory allows us to look a bit 'deeper' into the obtained results. We can not only obtain a prediction pointing to a specific type of outage but also levels of Pignistic probabilities. An illustration of results presented in such a way – the probabilities – is included in Table 5.5.

Table 5.5: Pignistic probability for various samples and the reference values, [144] ©2020 IEEE

Id	Reference	NoO	Wind	Snow	Icing
1	NoO	**0.9981**	0.0005	0.0008	0.0006
2	NoO	**0.9653**	0.0077	0.0093	0.0177
3	NoO	0.0579	**0.3232**	**0.2362**	**0.3828**
4	Wind	0.0047	**0.9756**	0.0098	0.0098
5	Wind	0.2342	**0.7181**	0.0246	0.0230
6	Wind	**0.8203**	0.1684	0.0056	0.0056
7	Snow	0.0286	0.0145	**0.9402**	0.0167
8	Snow	0.4527	0.0103	**0.525**	0.012
9	Snow	**0.6642**	0.0798	0.1836	0.0724
10	Icing	0.0007	0.0085	0.008	**0.9827**
11	Icing	0.0297	0.0339	0.0559	**0.8804**
12	Icing	0.0058	**0.6912**	0.1294	0.1736

Let us take a close look at the presented results. If we look at rows 1, 2, 4, 5, 7, 8, 10, and 11, we see examples of correct classification. The highest probability values are associated with proper predictions – the 'Reference' column versus columns 'NoO', 'Wind', 'Snow', and 'Icing'.

The interesting cases are shown in rows 3, 6, 9, and 12. They represent misclassified predictions. When we compare the probabilistic values, we notice that the values reveal an interesting insight into the prediction process. It seems there is no agreement between classifiers, and that means that a user can make a judgment on how much she trusts the obtained predictions. For example, the values in row 3 show almost equal probability between Wind, Snow and Ice, although the correct prediction would be NoO. In rows 6, 9, and 12, we see more considerable differences between probability values, yet they convey a diminished trust in the predictions.

The obtained results and their statistical analysis indicate that three prediction models: RF, SVC, and MLP are the best – yet indistinguishable from the performance point of view – prediction models. In the case of the *WoutPS*'s performance, the ensemble-based architecture that uses DST has provided the best results. Additionally, this architecture enables a user to look 'inside' the results and learn a bit more about a way how predictions are determined.

5.6 Conclusion

The ability to predict outages, and even their type, is of significant importance for power utilities. In this chapter, we present the result of a process of constructing a Weather Outage Prediction System *WoutPS*. Special attention has been put to the selection of the most suitable prediction model and the system's architecture. The obtained results and their statistical analysis indicate that three prediction models: RF, SVC, and MLP are the best – yet indistinguishable from the performance point of view – prediction models. In the case of the *WoutPS*'s performance, the ensemble-based architecture that uses Dempster-Shafer Theory has provided the best results. Additionally, this architecture enables a user to look 'inside' the results and learn a bit more about a way how predictions are determined.

Chapter 6

Knowledge Graph Representation of Power System

6.1 Introduction

A distribution level power system is a complex and connection intense structure. It can be characterized by the presence of multiple components of different type – many of them 'small' and of a low economical value when compared with elements of a transmission level system – that are connected in a chain-like mode. Such a nature of a distribution grid makes it quite tedious to oversee, analyze, and maintain.

A power utility keeps information about its grid in relational databases spread across multiple organizational units. In order to 'see' a part of the system, i.e., components, their types and parameters, and details of their physical connections, as well as information about linked customers and their locations, a number of operations related to accessing few databases and performing multiple operations on tables containing information about components of interest need to be performed.

Recently, a new type of databases that offer a different data representation format become more and more popular. They are graph-based databases that represent information as network of nodes and relations between them. Such databases provide a number of benefits:

- easiness of data integration from multiple data sources;

- easiness of expansion, i.e., ability to continuously append new pieces of data;

- simplicity of accessing individual pieces of information and performing queries, especially in the context of connections/relations between data nodes;
- openness to develop and execute variety of algorithms on graphs in order to gain additional information and enhancing graphs, for example, computing a length of sequence of specific types of nodes.

This chapter shows how the power grid's components can be stored in a format of the knowledge graph in a graph database, Neo4j. We provide details related to concepts and relations between them that constitute nodes and edges of a graph. We briefly describe a graph – called GridKG – representing a distribution grid. A few examples showing how such a database can be utilized are described.

The chapter's main contribution is constructing a knowledge graph that integrates information about the topology of the power system with meta data about equipment, customers and power outages. We enhanced the knowledge graph with the data generated by algorithms we developed to identify upstream and downstream devices, the number of customers connected to them, as well as the impact of power outages . This additional data leads to a holistic view of the system. All this would allow us to gain further insight into the system's characteristics while analyzing the grid.

6.2 Knowledge Graphs – Brief Overview

A knowledge graph is basically a data management system which combines various types of data and utilizes graphs to represent information and knowledge [146]. Initially introduced by Google, the knowledge graph concept was utilized to optimize search engine performance with the information gathered from various sources. Knowledge graphs such as BabelNet [147], DBpedia [148], WordNet [149], Microsoft's Probase [150] and Google Vault [151] are created focusing on text-based extraction of data from the web content. The existing generic knowledge graphs prove their usefulness in applications such as semantic search, and information fusion from various sources.

The Resource Description Framework (RDF) data format [46] introduced by the Semantic Web as a standard for Linked Open Data is a popular graph-based data format in which each piece of data is stored as a RDF triple that contains: two entities, two nodes in a graph, called a subject and an object; and a relation between them, an edge in a graph, called a property. Processing data represented as knowledge graphs in an RDF format is generating a lot of attention. There are multiple works focusing on different aspects related to RDF- based Knowledge Graphs: from their construction [47], via storage [48], querying strategies [49], [50] and extracting information [51], to applications [52], [53], just to mention a few. The fact that subjects of triples could be also objects of another triples, and vice versa, means that we deal with a network of entities highly interconnected via properties.

A single RDF-triple <subject-property-object> can be perceived as a feature of an entity identified by the subject. In other words, each single triple is treated as a feature of its subject. Multiple triples with the same subject constitute a description of a given entity.

Quite often a subject and object of one triple can be involved in multiple other triples, i.e., they can be objects or subjects of other triples. In such a case, multiple definitions can share features, or some of the features can be centres of other entity descriptions. All interconnected triples constitute a network of interleaving definitions of entities.

6.3 Related Work

A major challenge in managing the power grid with hundreds of thousands of power devices is how to collect, analyze, and manage the equipment of the grid [152], [153]. Using a knowledge graph can greatly enhance the efficiency of knowledge retrieval and greatly improve the quality and accuracy of the search results. However, few studies have focused on domain-specific applications of knowledge graphs in power grid operations in the power industry realm. In the following, a few examples of using graph-based data representations in power systems is provided.

Industrial utilization of knowledge graph at Siemens was an essential move toward intelligent engineering and production, improving workflow performance and data accessibility. The proposed knowledge graph could overcome traditional database challenges, provide an integrand view of data, improve search functionalities and data control [154]. Enterprise-level power equipment knowledge graph, improved power equipment management, and enhanced efficiency in retrieving, classifying, and updating relevant information [146]. In [155], an architecture for modeling and energization analysis of the IEEE 118-bus system topology is designed. It demonstrated the higher performance of the graph database compared to a relational database for the power grid analysis.

Various relation extractions between power grid asset entities have been investigated in some recent works. An ontology extraction framework was created to automatically build power terminal knowledge graphs [156]. The framework facilitates the sharing of heterogeneous data between multiple sources. A management system based on ontologies has been developed, capable of modeling interactions across multiple domains [157]. This study demonstrates the advantage of using ontologies to build decision support tools and provides an example of implementing optimal power flow in knowledge graphs. The AI-enhanced labeling method to create a knowledge graph was developed to group the power grid equipment with similar characteristics [153]. In [158], the authors demonstrated the knowledge graphs ability to evolve based on business needs and enhance the database expansion process and response time.

The deployment of advanced metering infrastructure and sensors has accelerated as smart grids grow, which generated unprecedented amounts of multi-source heterogeneous big data [159]. For a smooth transition to the next generation of smart grid systems, knowledge graphs assist in incorporating additional information and connectivity across all devices [153]. In smart grids, the information island problem was addressed using grid equipment knowledge graphs considering the multifaceted equipment nature and interequipment relationships [160]. Based on the knowledge fusion, [161] proposed a multi-source information fusion method to provide a unified knowledge base to utilize power equipment data efficiently. [162] utilize knowledge graphs to integrate power grid and environmental information and use it as a reference to perform signal correlation algorithms for abnormality detection in the grid. Dispatch knowledge graph for power grid was developed by extracting the entities and identifying the relationship patterns in dispatching behavior using natural language processing, and machine learning techniques [163]. These efforts and studies confirm the growing interest and demand in power grid knowledge graphs.

6.4 Power Knowledge Graph

In this section, we introduce **PowerLOV** – a Power Linked Open Vocabulary – that defines classes and relations that should be used to build a graph-based representation of any power system. The vocabulary is built as an extension of the ontology *PowerSystems.owl*. We present a list of concepts of **PowerLOV** as well as a set of relations. We follow up with a number of exemplary cases that illustrate, as well as emphasize, benefits of representing Power Distribution Grid as a knowledge graph.

6.4.1 Vocabulary Overview

As we have mentioned earlier, the proposed approach for representing a power system in a form of a knowledge graph requires defining a special vocabulary that allows to combine all three categories of information. In order to construct such a vocabulary, we have followed an example of the existing power system ontology called *PowerSystem.owl* [157]. Yet, to satisfy our needs, we have added more concepts and relations that address our particular requirement of representing diverse type of information.

6.4.2 Concepts

Some concepts that enable construction of power system model have already been defined in *PowerSystem.owl*. Yet, we have extended this ontology to accommodate concepts and categories required to represent such information as geographical locations, events, and maintenance activities. These concepts allow to build a view/representation of the system beyond its basic electrical components.

Two essential concepts – building blocks of any knowledge graph aimed at representing a power system – are *Element-Asset* and *cNode*. *Element-Asset* is equivalent to the category *ElectricalEquipment* from *PowerSystem.owl*. In **PowerLOV**, the class has the following attributes: *id* that is the same as the asset ID assigned to it by a utility, and two sets of coordinates x, y and *longitude* and *latitude* as identifiers of its geographical location. The second concept – *cNode* – is a fictitious connection point. Each *Element-Asset* is connected to two of them. They play the role of points linking two adjacent *Element-Asset*s.

Each *Element-Asset* – a node in the graph – is further described by connecting it to other nodes. In other words, each node is defined by a number of features, i.e., connections to nodes representing additional information related to *Element-Asset*. These nodes represent facts that define specifics of an asset. They are divided into three different groups: 1) nodes identifying a type of an asset and details of its specification; as well as information containing details of maintenance activities performed on an asset; 2) nodes related to system topology, i.e, their geographical locations; and 3) nodes describing system events in which assets were involved.

The first group is organized as a hierarchy of classes, Fig. 6.1. We defined the following superclasses:

- *Components* – it is a superclass that contains a number of subclasses called *ComponentType* that provide a means to identify a type of an asset; we have the following subclasses: *Switch*, *VoltageRegulator*, *EnergyMeter*, *ElectricalLine*, *PowerGenerator*, *CapacitorBank*, *SwitchGear*, *Sectionalizer*, *Elbow*, *Recloser*, *Fuse*, *SubstationBus*, *FaultIndicator*, *IsolationPoint*, *Breaker*, *SubstationTransformer*;

- *ConnectionStatus* – with subclasses *ConnectionStatusTypes* that indicate an operational status of an asset: *Connected* or *Opened*;
- *Phases* – it a superclass of *PhaseValue* that stands for phase(s) to which an asset is connected;
- *Orientations*– with two subclasses of a type *OrientationType* that indicate if an asset is *undegroundCable* or *overheadLine*;
- *Voltages* – with subclasses *VoltageValue* that specify rated and connected voltages of an asset;
- *Customers* – with a subclass *Customer* instances of which represented customers of the modelled system.

The second group includes classes used to identify specifics of geographical location of assets. Before we provide a short description of the classes, we would like to indicate that besides such obvious location indictors as service area and feeder, to which an asset is (indirectly) connected, we have introduced a much finer division of service areas – a grid of polygons. The motivation to introduce a grid of polygons has been twofold: 1) to better localize different elements and events at the resolution that is 'between' service areas and geographical coordinates; and 2) to prepare an introduction of additional data – for example, weather – that enhances even further possibilities of analysis of power system behaviour at different (weather) conditions. The classes of the second group are:

- *ServiceAreas* – with a number of subclasses *ServiceAreaNames* representing names of the utility's service coverage zones;
- *Feeders* – it is a superclass of *FeederID* that provides identification of an upstream feeder to which an asset is (indirectly) connected;

- *GridCell* – identification of an individual grid cell in the created grid of polygons.

In the case of the third group, we define two categories that allows us to provide details regarding events, for example outages, that occurred in the modelled system and effected an asset:

- *OutageEvent* – provides details regarding an outage, such as time of event, power interruption interval, supplementary cause;
- *OutageCauses* – with a subclass *OutageCauseType* that identifies a possible cause of an outage, examples of recognized causes are: *Lightning*, *DefectiveEquipment*, *ForeignInterference*, *AdverseWeather*, just to name a few.

6.4.3 Relations

As we have mentioned earlier, nodes of a knowledge graph are connected via well-defined and semantically meaningful links – relations. These relations allows for characterizing assets and providing details describing assets. In the context of **PowerKG**, we recognize a number of relations:

- *Connection* – links *Element-Asset* to *cNode*;
- *hasComponentType* – links *Element-Asset* to *ComponentType*;
- *hasConnectionStatusType* – links *Element-Asset* to *ConnectionStatusType*;
- *hasCustomer* – links *Element-Asset* to *Cusomer*;
- *hasFeederId* – links *Element-Asset* to *FeederId*;
- *hasGridCellId* – links *Element-Asset* to *GridCellId*;
- *hasLocationType* – links *Element-Asset* to *LocationType*
- *hasOrientationType* – links *Element-Asset* to *OrientationType*
- *hasOutageEvent* – links *Element-Asset* to *OutageEvent*

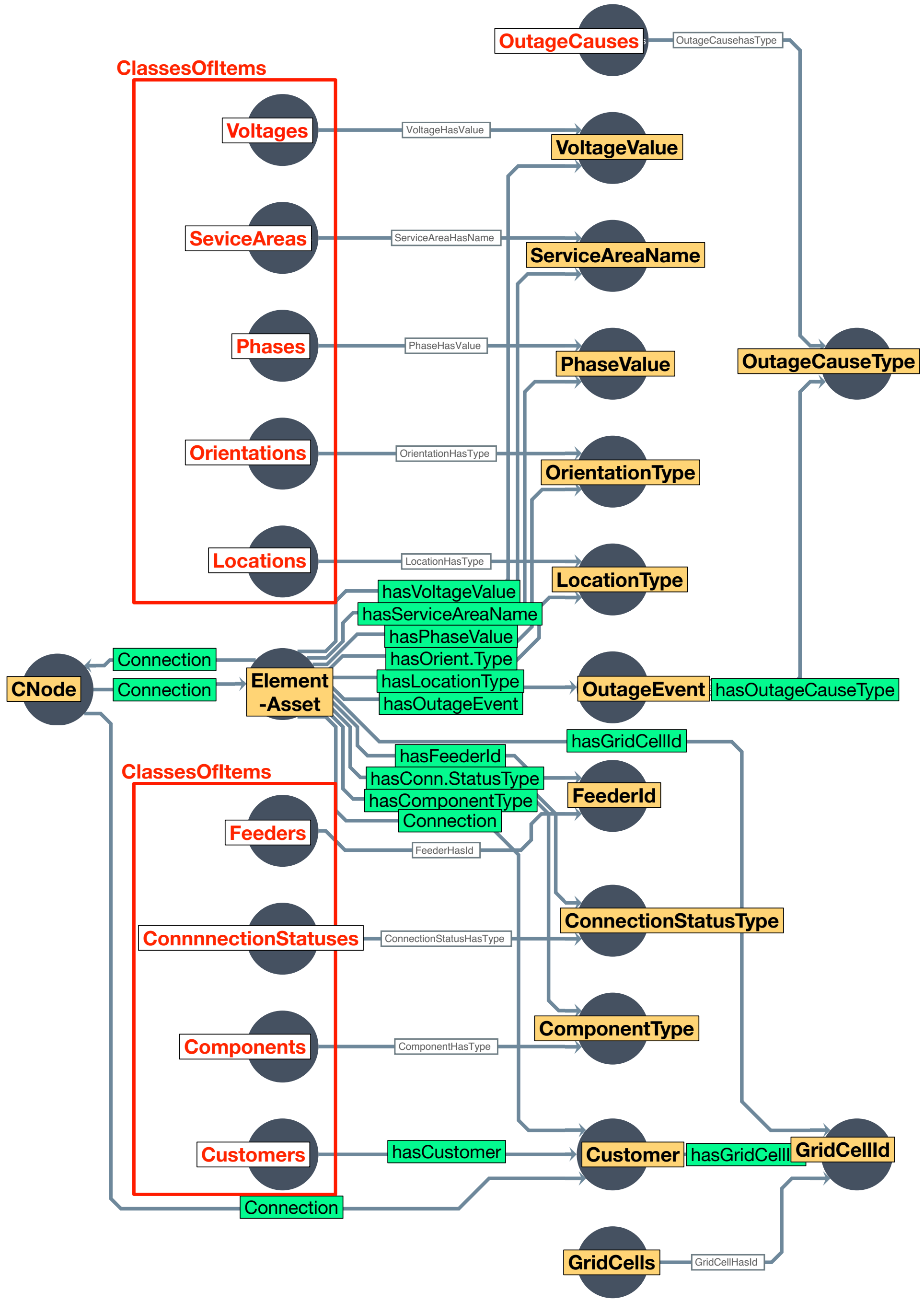


Figure 6.1: Schema of **PowerKG**: red names represent superclasses of items

- *hasOutageCauseType* – links *OutageEvent* to *OutageCauseType*
- *hasPhaseValue* – links *Element-Asset* to *PhaseValue*;
- *hasServiceAreaName* – links *Element-Asset*to *ServiceAreaName*;

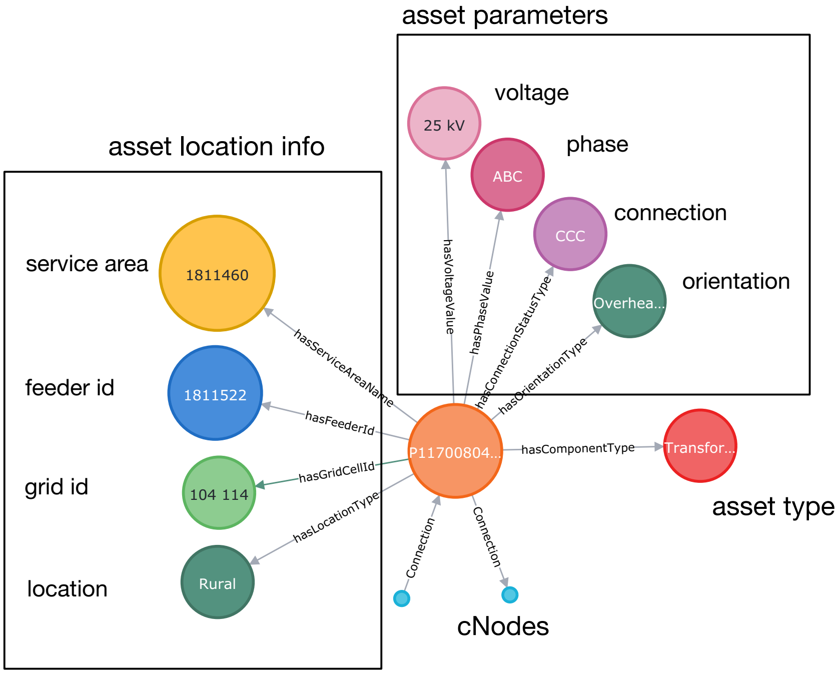


Figure 6.2: RDF triples with *Element-Asset* as their subject

- *hasVoltageValue* – links *Element-Asset* to *VoltageValue*;

All concepts and relations created for **PowerKG** are presented in Fig. 6.1: superclasses are in red, subclasses and both *Element-Asset* and *CNode* are in yellow, while relations used to define details of *Element-Asset* are in green.

6.4.4 Simple Examples

Before we present more involving examples, we include a simple relation-based description of of an asset, Fig. 6.2. It is a transformer, in the service area *1811460*, in the polygon grid cell *104-114*, connected to the feeder *1811522*, at the *Rural* location. Its rating is *25kV*, it is connected to all phases (*CCC*) *ABC*, and its orientation is *Overhead*.

As we have mentioned earlier, the *Element-Asset* in Fig. 6.2 is an example of a basic building block of graph representing a power system. Each *Element-Asset* that is an electrical component is linked – via the relations *Connection* – to *CNodes*. Such an arrangement links all components together 'electrically'. Links to other, non-electrical, nodes determine details of components. A fragment of a graph-represented system that contains five *Element-Asset*: *Fuse*, *Transformer* and three *ElectricalLines* is illustrated in Fig. 6.3.

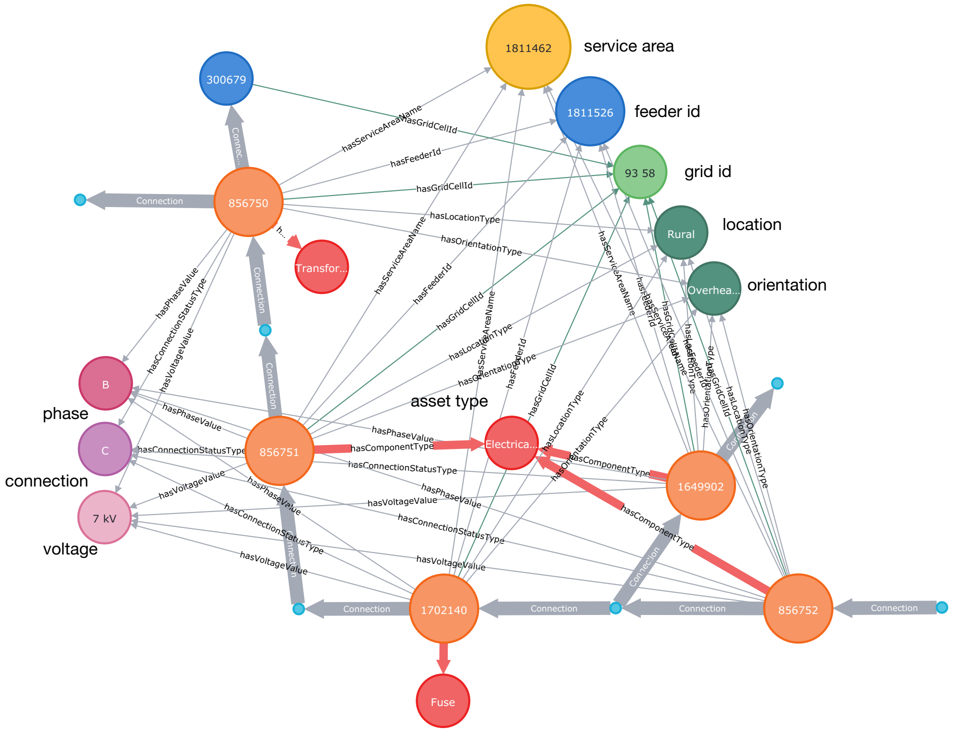


Figure 6.3: A fragment of the distribution grid with a transformer, three electrical lines and a fuse: electrical connections are shown as thick grey lines.

6.5 *GridKG* – Graph-based Representation of Distribution Grid

The defined above vocabulary **PowerLOV** is used to construct a representation of real power distribution grid. A graph-based model of a full-sized power distribution system is called, hereafter, **GridKG**. It contains more than 2 millions of nodes and 10 millions of connections. Fig. 6.3 shows its very small fragment.

GridKG combines information of very different nature: details of electrical equipment, system topology, and system events. All pieces of information are semantically interconnected. The connections between nodes of **GridKG** create an opportunity to process available data and analyze the modelled grid in multiple different ways. The following two sections – this one and Section 6.6 – provides a number of examples what can be done.

6.5.1 Electrical Paths

One of the most interesting examples of processing data included in **GridKG**, and enhancing it at the same time, is identification of electrical paths in a system. This allow for determining upstream and downstream elements, and as the consequence analyze the graph from yet different points of view.

In a graph-based representation of a grid, an electrical path is considered as a sequence of tuples

$$\langle \mathit{cNode}_x \rightarrow c \rightarrow \mathit{Element\text{-}Asset}_p \rightarrow c \rightarrow \mathit{cNode}_y \rangle$$

where c represents the relation *Connection*. This tuple is composed of two triples connected via the same $\mathit{Element\text{-}Asset}_p$, i.e., $\langle \mathit{cNode}_x \rightarrow c \rightarrow \mathit{Element\text{-}Asset}_p \rangle$ & $\langle \mathit{Element\text{-}Asset}_p \rightarrow c \rightarrow \mathit{cNode}_y \rangle$. An example of an electrical path is shown in Fig. 6.3 where nodes representing *Element-Asset*'s are connected with via thick grey links. The proposed process of 'building' paths is equivalent to 'stitching' together a number of tuples in a way that cNode_y of the predecessor tuple matches cNode_x of the successor tuple. Algorithm 2 shows the simplified steps.

An input to the algorithm is a queue **Q** of triples ⟨*Element-Asset*-c-*cNode*⟩, where *Element-Asset* is of type *Breaker*. Based on the triples from the queue **Q**, the Breadth-First search algorithm finds a sequence of adjacent *Element-Asset*'s. If it happens that the encountered *Element-Asset* is of type *Switch* then its *ConnectionStatusType* has to be *Connected*. If it is *Open* it is not considered as a part of a continuous electrical path.

We define a special attribute of the *Element-Asset* – called *level* – that is used to 'keep' track of the elements' position in the path. For each *Element-Asset*, the algorithm modifies a direction of connections from an element of a lower value of *level* to the adjacent element with a higher value of *level*. The algorithm finishes when the queue **Q** is empty.

A process of establishing electrical paths allows us to easily identify upstream and downstream elements. This can be used, for example, to determine a number of customers that 'depend' on each element. The deployment of a single algorithm augments each element with an additional information,

Algorithm 2 Multi-Source Path Search, [152] ©2020 IEEE

```
Input:
    $\mathbf{Q} \leftarrow$ list  of  triples $\langle cNode - c - Breaker \rangle$

Initialization:
  $level = 0$
  set: nextElements = {},                                   ▷ Elements to visit
  set: S ={},                                               ▷ visited Elements

while Q is not empty do
    $size \leftarrow$ numberOfTriples(Q)
    for  1 to $size$ do
        $\langle cNode - c - Element \rangle \leftarrow$ getTriple(Q)
        if id($Element$) $\in$  S  then                         ▷ Element visited
            continue
        level of $Element \leftarrow level$
        $\mathbf{S} \leftarrow Element$
        $cNode\_next \leftarrow$ getNextcNode($Element$)
        nextElements $\leftarrow$ getNextElment($cNode\_next$)
        for each $Element\_next \in$ nextElements do
            setCDirection($Element$, $cNode\_next$, $Element\_next$)
            if  not (type($Element\_next$) == Switch &
                  state($Element\_next$) == Opened) then
                $\mathbf{Q} \leftarrow \langle cNote\_next - c - Element\_next \rangle$
            end_if
        end_for
    end_for
    $level \leftarrow level + 1$
end_while

return
```

stored in yet another additional attribute *numCustomer*, that indicates how many customers are located downstream of the element.

6.5.2 Review of Switching Elements

With electrical paths identified, we can utilize **GridKG** to learn more about connections between different *Element-Asset*s. One of possible ways of learning more about the grid is finding out all downstream (and upstream) elements and connections from a given element of the system.

Example A

Let us ask **GridKG** about downstream paths/components from a specific location. Additionally, we can impose a condition regarding a maximum number of components composing the path, and retrieved lengths of these paths together with some basic information about events – outages in our case – involving the paths' components.

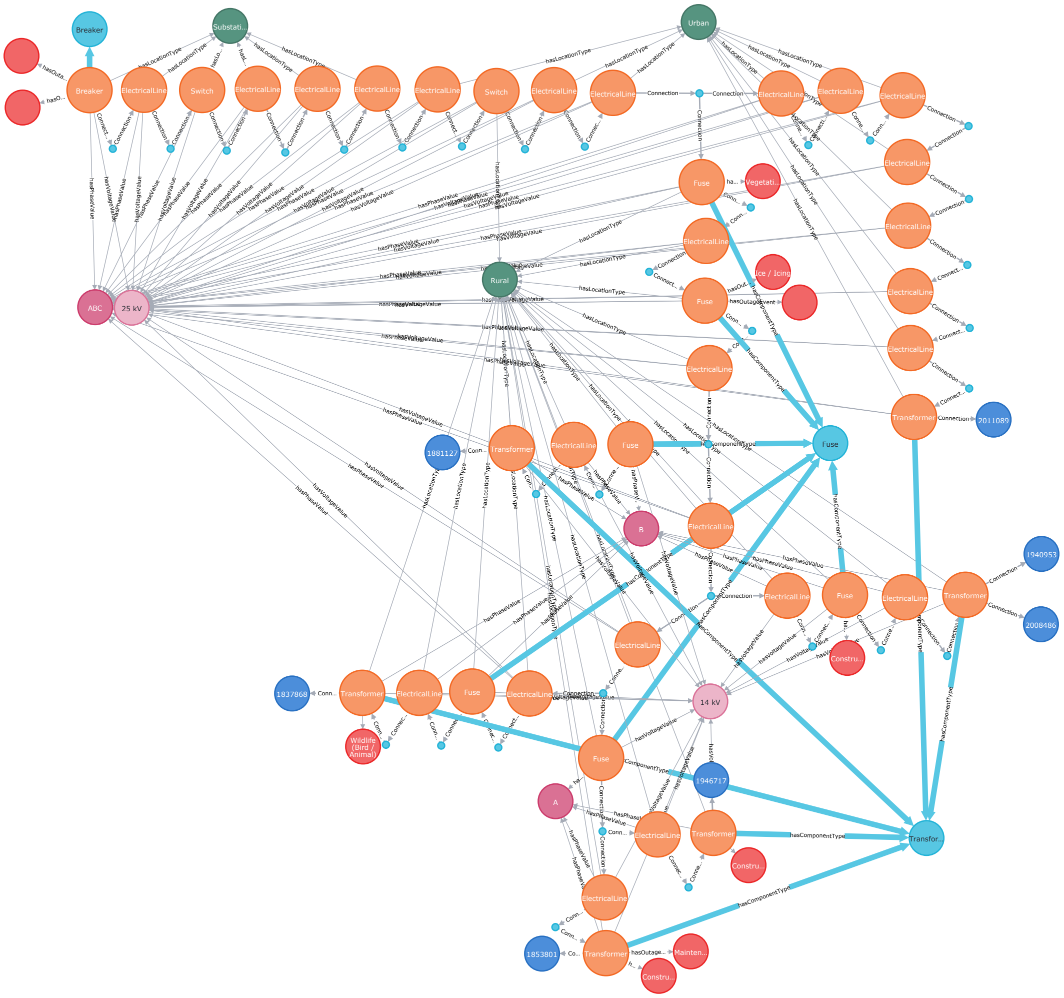


Figure 6.4: *Elements* of downstream paths: their types identified by links to light blue circles where each circle represents a different type of electrical component; their voltages – light pink circles; their phases – dark pink circles; as well as outages – red circles.

The *Neo4j* query, in the language *Cypher*, is shown below. We provide a starting element *source*, and a number representing a maximum number of components. The query returns all downstream connected elements, Fig. 6.4. Additionally, it returns lengths of the electrical paths – there are seven paths of the length: 465.21m, 445.09m, 595.25m, 595.25m, 1272.90m, 1325.93m, and 1599.44m, respectively. It also returns the Element-Assets on the path associated with power outages. In Fig. 6.4, we see information about types of elements, their connected voltage values and phase, as well as outage events (for clarity, some details are not shown).

```
MATCH path = (source:Element{mslink:7892771})
               -[:Connection*..40]
              ->(c:Customer)
RETURN path, reduce
              (total = 0, e IN nodes(path) |
        CASE
            WHEN e.LengthValue IS NOT NULL
            THEN total + e.LengthValue
            ELSE total
        END )
AS totalLength, [element IN nodes(path)
WHERE (element) -[:hasOutageEvent]-(:OutageEvent)]
```

Example B

Let us take a look at another way of analyzing the distribution system, i.e., finding a number of protective devices that exist on a path upstream from a specific location.

This time, the query – presented below – returns a path of upstream elements from the specified *Element-Asset* up to the breaker, Fig. 6.5. As we can see, the path includes two *Element-Asset* of type *Switch* – this allows us to locate the first and the second of these devices on the path. Additionally, we obtain more information about the path itself – its components, and information about them, as well as a number of customers connected downstream from the specified *Element-Asset* – it is 1828 in the case of the system represented by **GridKG**. This query can be used to investigate how many more customers will be affected by the activation of backup protection in case the primary protection fails to act. It is possible to investigate the whole grid and find out the critical locations requiring closer attention to its protection system design.

```
MATCH  path = (source:Element{mslink:1169863288})
              <-[:Connection*]
              -(protective:Element)
WHERE  protective.ComponentType IN ['Breaker','Switch','Fuse']
RETURN protective.NumCustomer, path
LIMIT 10
```

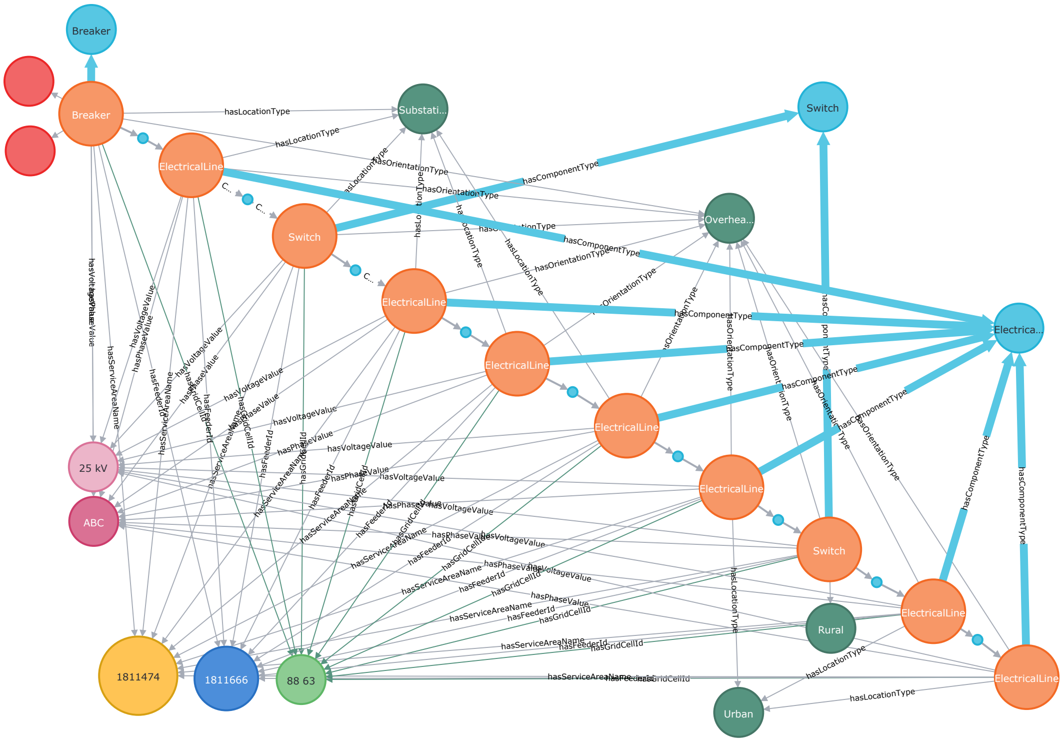


Figure 6.5: Location of Primary Switch/protective devices on the upstream path from top-left element to the primary breaker (down-right corner).

6.5.3 Review of System Events

Integration of different types of information in **GridKG** enables analysis of system components, in particular their status and maintenance activities, as well as system events occurring at specific locations.

Example C

Let us retrieve from **GridKG** a list of outage events that occurred downstream from a specific *Element-Asset*. We want to know details of the outages as well as a graph representing a fragment of the system with information about the components. For illustrative purposes we have limited the response to 10 outages. The query we use is shown below.

```
MATCH path = (source:Element{mslink:7892771})-[c:Connection*..]->(
    element:Element) -[:hasOutageEvent]->(outage:OutageEvent)
RETURN [outage.OutageCause, outage.OutageTime,element.NumCustomer,
     element.GridCellId, element.Longitude, element.Latitude] AS
    OutageInfo, path
```

```
LIMIT 10
```

The obtained data in a form of a list of outages is shown in Table 6.1. As we can see, details related to times and causes of outages are retrieved. Additional information, types of involved *Element-Assets* and their electrical specifications, as well as details related to effected customers are included in a graphical view of the obtained data, Fig. 6.6. The figure shows that one of the involved elements was *Fuse*, connected to the phase A of $14kV$ line, that was involved in seven outages. Three of them were caused by a wildlife, and four were of an unknown cause. Other outages were associated with a transformer, a switch and another fuse. Via interacting with the *GridKG* more information about specific outages, involved *Elements-Asset*s, as well as the effected customers can be easily obtained and examined.

Table 6.1: Details of 10 outages (selected information)

Id	**Time**	**Primary Cause**	**Secondary Cause**
1	1997-12-07 14:23	Unknown	–
2	1997-06-24 23:59	Defective Equip.	Electrical Failure
3	1997-06-28 22:04	Unknown	–
4	1998-05-13 06:06	Unknown	–
5	1999-09-10 18:08	Unknown	–
6	2000-08-14 10:53	Foreign Intrf.	Wildlife (bird/animal)
7	1998-07-27 18:38	Foreign Intrf.	Wildlife (bird/animal)
8	1998-08-09 02:25	Unknown	–
9	1996-07-30 08:15	Foreign Intrf.	Wildlife (bird/animal)
10	1997-06-28 21:53	Unknown	–

Example D

Another query could be related to outage events occurring in a specific service area and linked to a specific *Element-Asset* that satisfies particular conditions.

For example, we can ask for outages linked with *Switch*es that went through maintenance activities taking place during a specific period of time. An additional condition could be related to finding the switches on the path that has at least fifty customers downstream.

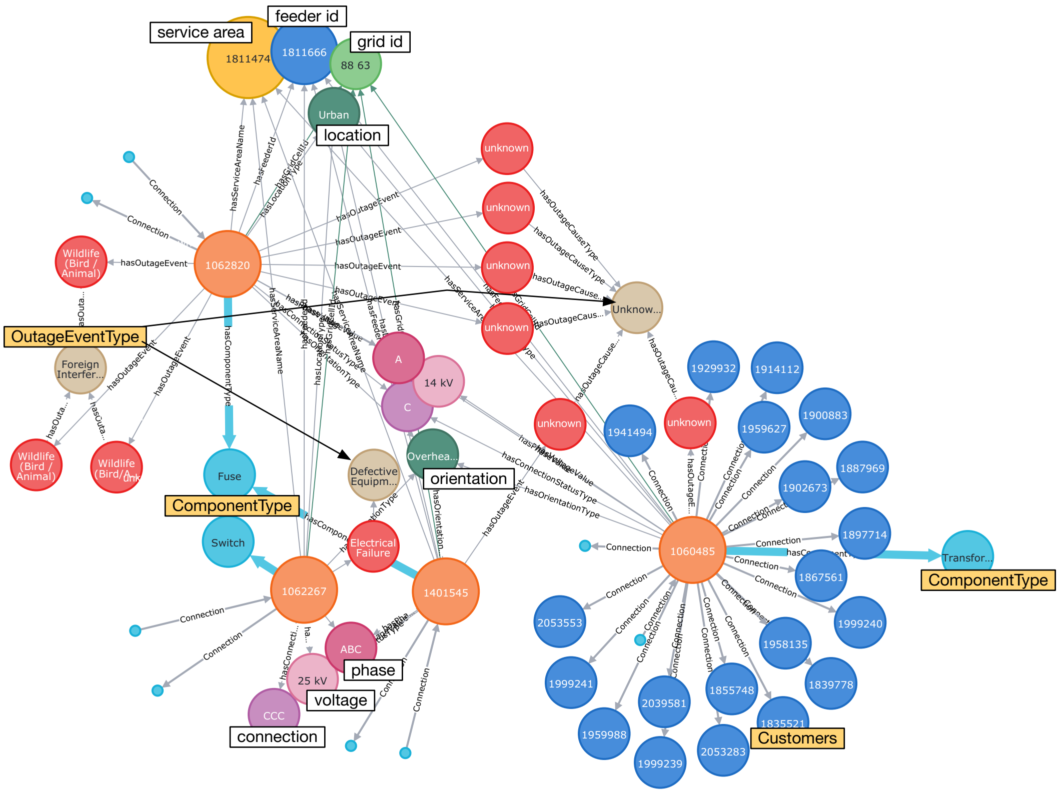


Figure 6.6: Details of 10 outage events downstream of a specific *Element-Asset*.

The *Cypehr* query is shown below. As we can see, we are asked for a list of switches involved in the outage caused by *Adverse Weather* and a maintenance activity *XYZ* done on the switches between *time1* and *time2*.

```
MATCH (switch:Element{ComponentType: ”Switch”, ServiceAreaName: ”
    St. Paul”})
        -[:hasOutageEvent]
        -(oe:OutageEvent{OutageCause : 'Adverse Weather'})
MATCH (switch)-[c:Connection*..]->(downstream_element:Element)
WHERE  switch.NumCustomer > 50
AND downstream_element.MaintenaceTime > time1
AND downstream_element.MaintenaceTime < time2
AND downstream_element.maintenaceType = ”XYZ”
RETURN switch, downstream_element
```

6.6 Analysis of Outage Severity

The disruption of energy delivery caused by power outages occurs quite often in the power distribution grid [144]. They result in high costs for power utilities, with an estimated $20 to $55 billion annual cost to the U.S. economy due to storm-related outages [1]. Predicting the severity and impact of power outages help power utilities to plan ahead for resource allocation which will lead to fast recovery after power outages, improve customer satisfaction, grid reliability, and profitability of utilities [144].

As reported in the literature, analysis of power outages focuses on predicting a number of outages [26], [28], [30]–[32].

Generally, the effort is put into predicting the number of outages due to adverse weather conditions as per single cell of a grid that a specific area had been divided into. Yet, it seems that a more appropriate way to determine the effect of outages is to consider the number of affected customers. Some published work suggests predicting the number of affected customers instead of the number of outages [164]. This would provide the ability to estimate the effect of adverse weather conditions on the power system reliability by calculating the SAIFI [1] index. But, development of such models is not easy. The models would require a large number of data points that often is not possible.

Estimation of power outage severity impact can be done by applying **GridKG** for estimating the number of affected customers. The graph provides an easy access to the power system topology. This alone allows to consider differences in locations between power outages and customers. An extreme weather event that leads to a power outage in one grid cell can affect customers in various grid cells positioned downstream to the power outage or the protective equipment locations. Therefore, information stored in the presented **GridKG** can help to better estimate the number of affected customers regardless of their physical location.

6.6.1 Monte Carlo Simulation: Overview

In order to illustrate an application of **GridKG** to a more demanding/complex analysis – in our case, the analysis of severity of outages – we have developed a

[1] The System Average Interruption Frequency Index (SAIFI), and defined as the average number of interruptions that a customer would experience.

process, based on **GridKG**, to estimate a number of affected customers. It applies a Monte Carlo simulation to construct distributions of affected customers. A pseudocode of the process in presented in the form of the algorithm 3.

Algorithm 3 Power Outage Severity Impact Simulation

```
procedure PREPROCESS(GridKG)

    Initialization:
        $k = 0.1$,                                              ▷ smoothing factor
        set: customers = {},                                    ▷ set of unique customers

    for each protect_equip$_i$ in $gridcell_{m,n}$ do
        $P_i = (N_i^{outage} + k)/(kN_{equip} + \sum_{i=1}^{N_{equip}} N_i^{outage})$    ▷ smoothed probability of power outages
        set: microgridCustomers = {}          ▷ set of unique customers supplied by a microgrid
        for each pcc in downstream of protect_equip do
            microgridCustomers ← getDownstreamCustomers(pcc)
        end_for
        customers ← getDownstreamCustomers(protect_equip)
        AffectedCustomers of protect_equip ← customers - microgridCustomers
            ▷ Subtract the microgrid Customers from all customers downstream of protective equipment to find the affected customers
    end_for

procedure SIMULATION(GridKG)

    Initialization:
        $S = 1000$,                                             ▷ number of simulation
        list: numberOfCustomers = [],  ▷ list of number of affected customers for each iteration of simulation
    for $i = 1, ..., S$ do                                      ▷ $S$ times random sampling
        select $n$ protective equipment according to the power outage probability distribution
        set: eventCustomers = {}    ▷ set of unique affected customers with an event with multiple outages
        for each selected protect_equip do
            eventCustomers ← getAffectedCustomers(protect_equip)
        end_for
        numberOfCustomers ← getLength(eventCustomers)   ▷ append to the list
    end_for

    return numberOfCustomers
```

In a nutshell, the idea behind the process is as follows: for a given number of required outages N in a given polygon P, we randomly select N outage locations – by an outage location we understand a location of a protective device, such as breaker, switch, and fuse – and identify how many customers are affected. This is repeated S times.

We consider two approaches to identifying locations of outages.:

- a location is picked based on uniform probability among all protective devices in given polygon;
- a location is picked based on a probability distribution calculated for a given protective device based on the involvement of this device in the previous (historical) outages information of which we have stored in **GridKG**; if, historically, a specific location and equipment are more prone to experience a power outage, there is a higher probability of being selected as in the simulation process.

For the example included in the chapter, we selected an arbitrary service area. As it has been mentioned earlier, the whole area has been divided into a grid of 10 × 10 km cells. The Figure 6.7 shows the studied region. The locations of historical power outages are depicted with the black dots. The power outage severity impact simulation is calculated for the blue grid cell as an example.

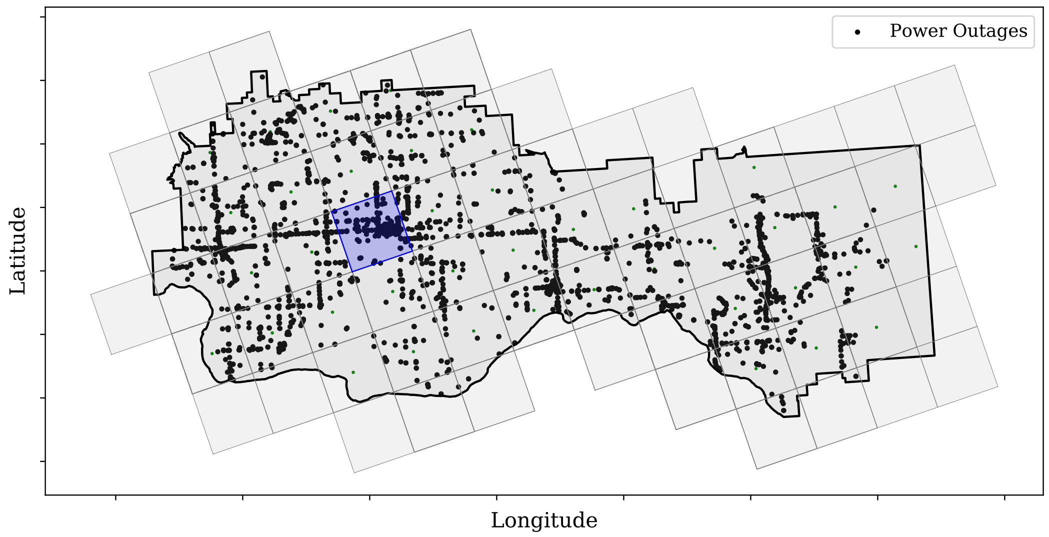


Figure 6.7: Service area grid cell with power outage locations - blue grid cell represents the cell that power outages in the simulation are located

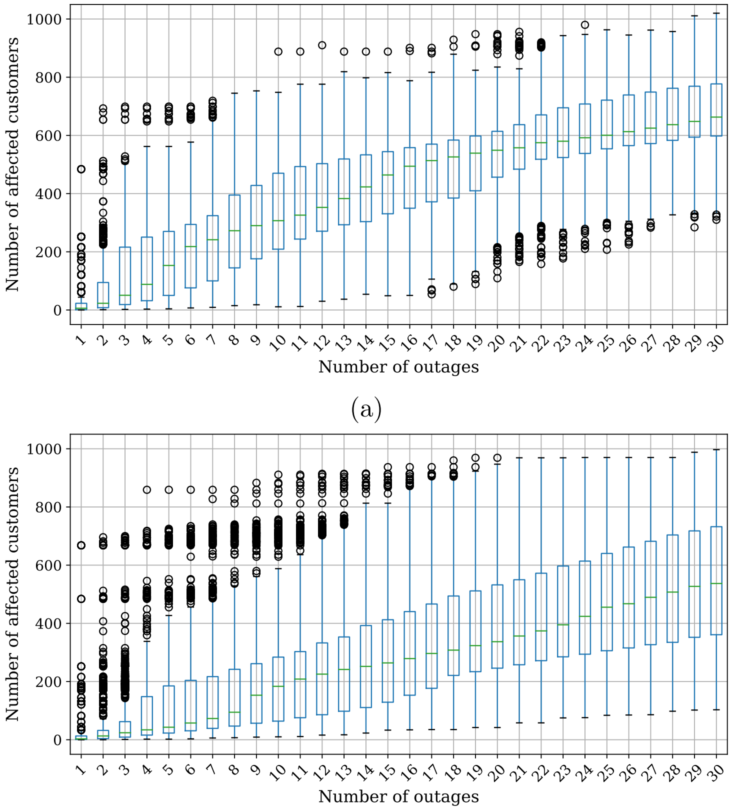
6.6.2 Analysis of Simulation Results

Various number of power outages, $N = 1 \ldots 30$, are considered to happen in the blue grid cell, Figure 6.7. Each of them is simulated $S = 1000$ times. We present the results in the form of box plots which show the distribution of the number of affected customers with respect to to the number of power outages in the blue grid cell, Figure 6.8. Each box depicts the first and third quarterlies of the data with green line representing the median of the data. The whiskers extend from both sides with the maximum length of 1.5 interquartile range (IQR) and represent the rest of the distribution, with small circles considered as the outliers.

Let us take a closer look, Figure 6.8 (a) shows the results when the historical information, stored in **GridKG**, is taken into account when 'selecting' an element associated with an outage; while (b) assumes uniform probability of selecting an element. In general, it can be seen that the slope of the line connecting the median values of the affected customers decrease for higher number of power outages due to having higher number of overlapping between affected customers. Yet, it looks that the inclusion of the historical data leads to a quicker rise in the number of affected customers yet a more consistent numbers. It can be said, based on that, the equipment that is more prone to be associated with the outages is 'responsible' for affecting more customers and should be given more attention.

It is interesting to compare the distributions of number of affected customers for both scenarios of selecting components involved in the outages. Figure 6.9 shows the distributions for $N = 15$ power outages. Once again, quite a difference can be observed. It reconfirms an earlier observation of importance of putting attention to the components heavily involved in outages in the past.

So far, we consider the number of affected outages without looking at the locations of customers. In this study, we look closer at the spatial distribution of outages. Figure 6.10 shows the studied service area with clearly marked polygons – green cells – with customers affected by the outages for the scenario

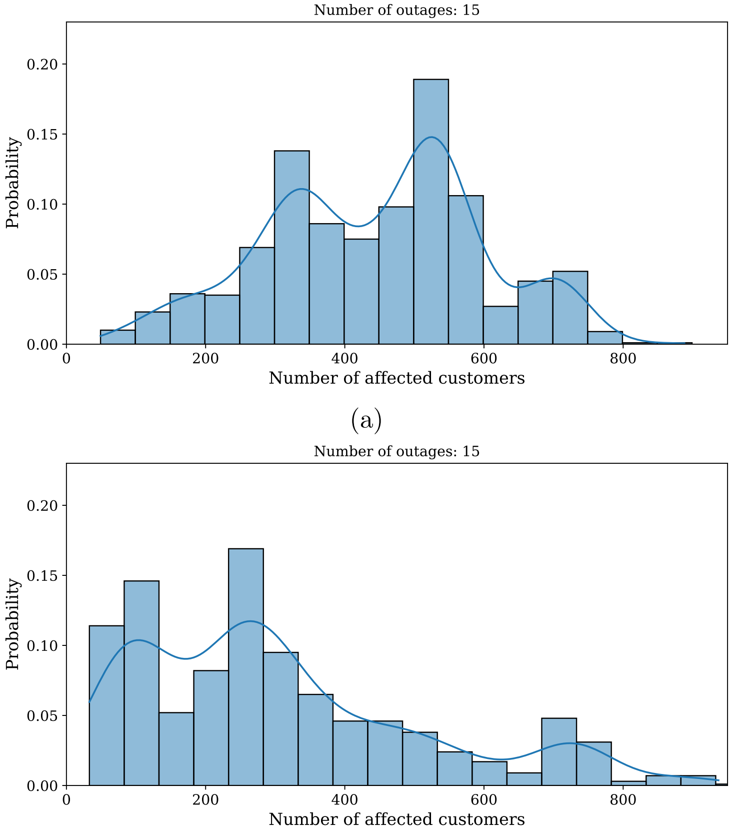


(a)

(b)

Figure 6.8: Number of all affected customers distribution for various number of power outages in the blue grid cell – equipment selection based on: (a) historical power outage probability; (b) uniform probability

when the involvement of elements is determined by the historical information. The darker green represents the higher number of affected customers. It shows how spatially areas are connected and how an outage in one specific area can affect the downstream customers. Yet, another view at this can be provided by looking at the distribution of the affected customers in each of the polygons. Such distributions are presented in Figure 6.11. It can be observed that a number of cells – 3, 6, 7, 8 and 9 – have 'binary' distributions. Among them, for the case of cell 7, none or all customers are affected. For the cells 3, 6, 8, and 9 it seems that only a fraction of customers are affected with the probability 0.6-0.75 if the outages 'reach' these areas.



(a)

(b)

Figure 6.9: Number of all affected customers distribution for $N = 15$ power outages: (a) equipment selection based on historical power outage probability; (b) uniform probability equipment selection.

6.6.3 Analysis of System with Micro-grids

Microgrids are small-scale energy systems powered by distributed energy resources and rely on them to work in a standalone islanded or grid-connected mode [165], [166]. The set of studies we present in this section focuses on effect of microgrid on the number of affected customers. We consider four microgrids connected to the grid. In the case of a power outage, microgrids disconnect themselves from the main grid at the PCC point and operate in islanded mode. Therefore, microgrids can lead to the reduction of the number of affected customers.

To incorporate micgrods in the simulations, the PCC points are added to the specific nodes of the **GridKG**, and are taken in consideration during calculations of affected customers.

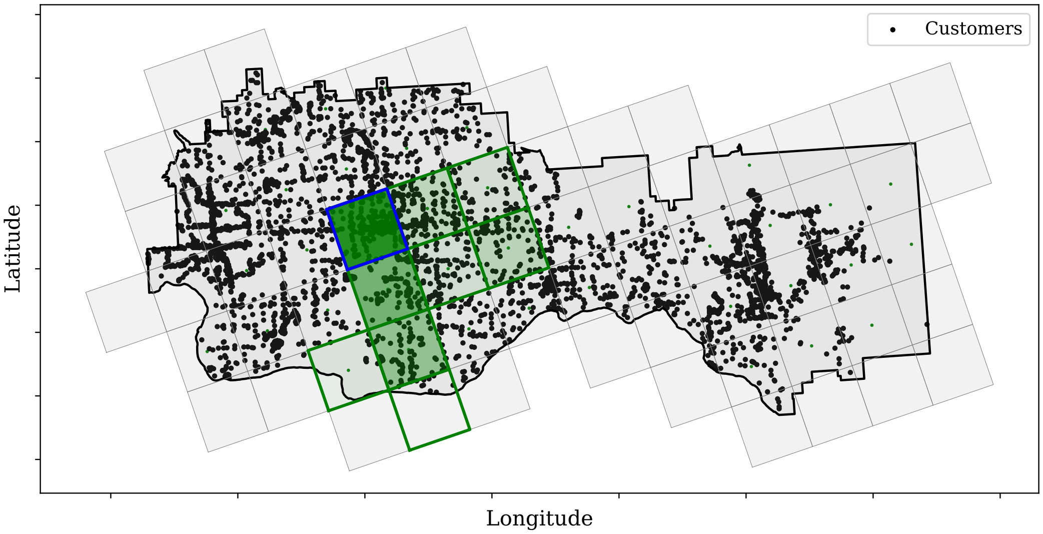


Figure 6.10: Service area grid cell and customers location - the green grid cell represents the customers that experience power interruption due to power outages in the blue grid cell - the darker colour represents the higher number of affected customers

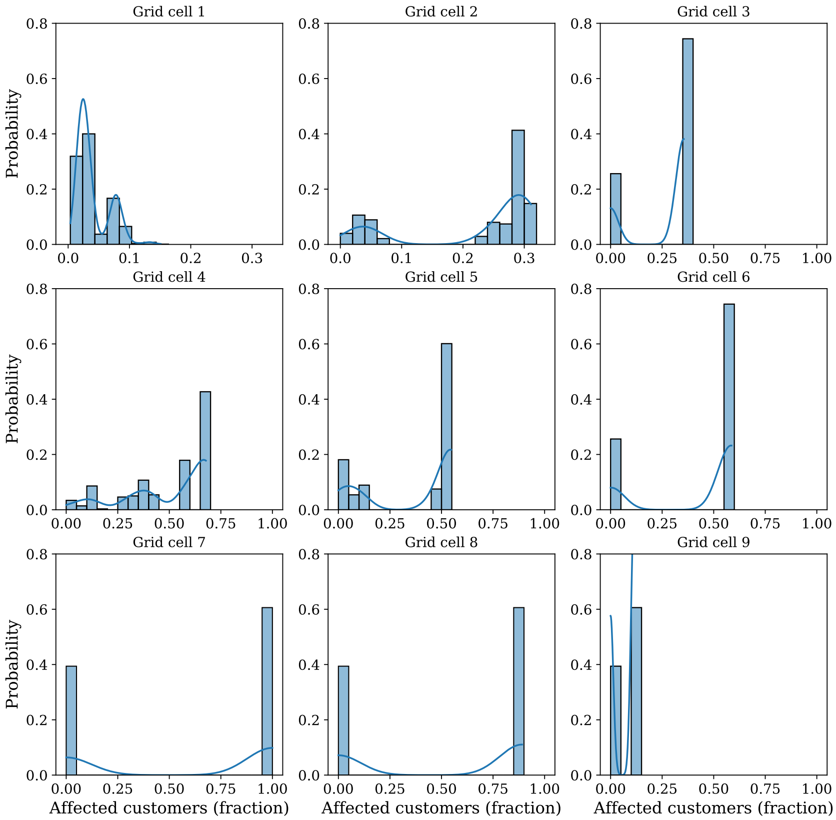
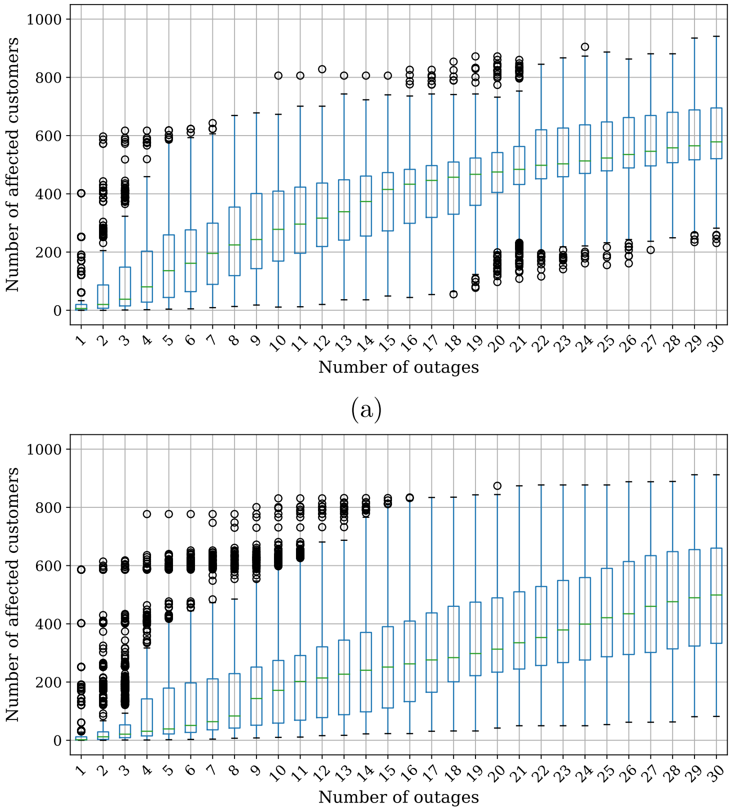


Figure 6.11: Number of affected customers distribution in the affected grids (green cells, Figure 6.10) due to power outages in the blue grid – equipment selection based on historical power outage probability.

As before, we provide two sets of plots: for a range of outages $N = 1 \ldots 30$, Figure 6.12, and for $N = 15$, Figure 6.13. The first figure shows the boxplot of number of affected customers for various number of power outages with the presence of the micrograms. As we can see, in comparison with plots in Figure 6.8, overall number of affected customers dropped as expected. The differences between outages involving elements historically more prone to outages versus uniform distribution of outages across all protective devices the similar as before. For the case of $N = 15$ outages, Figure 6.13, probabilities have similar trends.

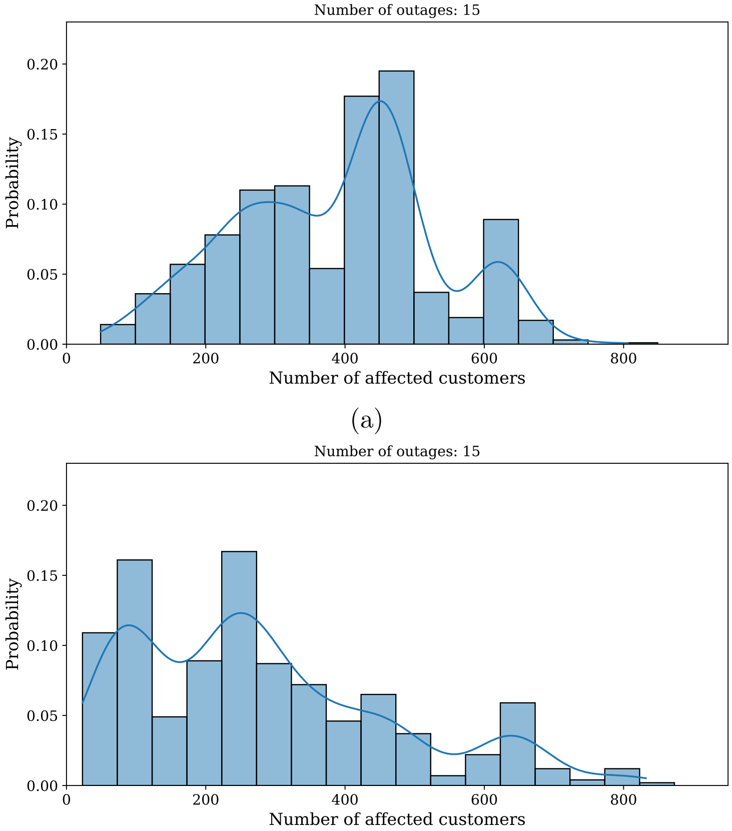


(a)

(b)

Figure 6.12: Number of all affected customers distribution for various number of power outages in the blue grid cell, considering the micro-grids connection to the PCC points: (a) equipment selection based on historical power outage probability; (b) uniform probability equipment

We also provide a set of plots allowing us to look inside multiple cells that contain affected customers geographically distributed, Figure 6.14. A visual inspection shows the largest different for the cell #5. The details regarding



(a)

(b)

Figure 6.13: Number of all affected customers distribution for 15 power outages in the blue grid cell considering the micro-grids connection to the PCC points: (a) equipment selection based on historical power outage probability; (b) uniform probability equipment

mean values of affected customers for each cell are in Table 6.2.

6.7 Conclusion

Today's distribution grids are complex networks constituted of multiple components. Power utilities collect and store, in relational databases, large amount of information about the grids' elements from transformers to individual poles. It is important for them to be able to have quick access the data describing components, as well as connections and relations between them.

We propose to use knowledge graphs as a suitable format for representing grid data. We describe some of the categories of nodes designed for representing different electrical elements and conceptual information describing those elements. We also define a number of relations between elements/concepts

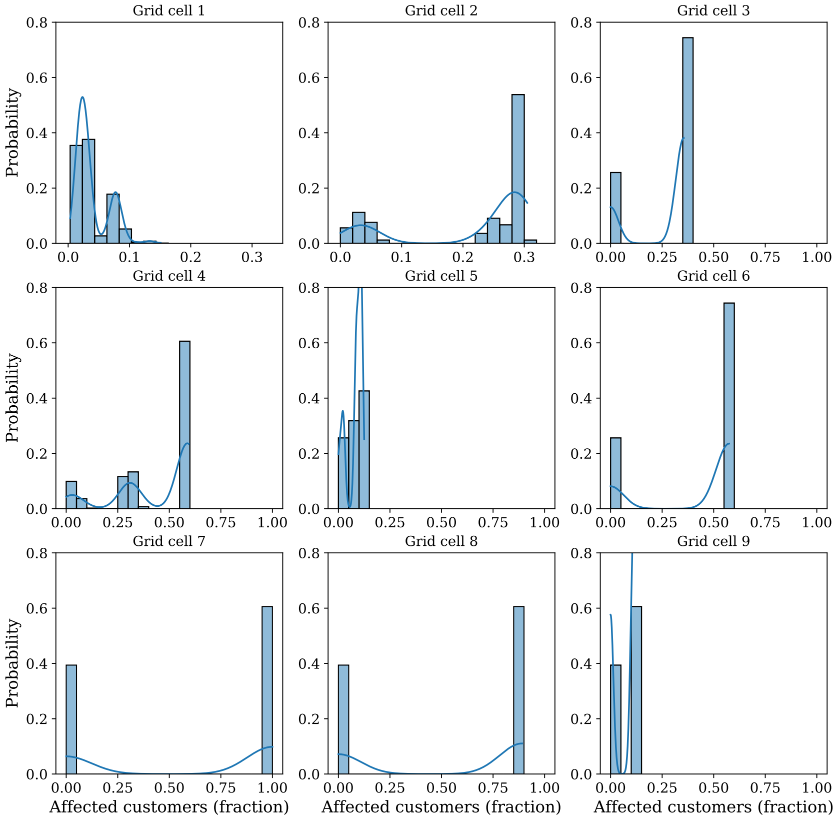


Figure 6.14: Number of affected customers distribution in the affected grids (green cells) due to power outages in the blue grid cell considering the micro-grids connection to the PCC points - equipment selection based on historical power outage

Table 6.2: Mean number of affected customers with 15 power outages in the blue gridcell

Location	Historical	Uniform	Historical w Micro-grid	Uniform w Micro-grid
All	442.83	305.04	385.81	281.14
Cell #1	121.70	174.98	117.9	170.14
Cell #2	36.38	13.18	35.53	12.84
Cell #3	16.36	5.58	16.36	5.58
Cell #4	108.48	47.65	95.45	42.82
Cell #5	**48.73**	**17.22**	**10.06**	**3.56**
Cell #6	37.87	12.93	37.2	12.7
Cell #7	8.48	3.87	8.48	3.87
Cell #8	64.23	29.36	64.23	29.36
Cell #9	0.60	0.27	0.60	0.27

that are linked to edges connecting nodes of the graph.

Finally, we illustrate utilization of a distribution grid knowledge graph. We propose an algorithm for identifying electrical paths in the grid. Further, we include a few graph queries that take advantage of the identified paths and allow for: determining a length of downstream path from a specific element; determining a sequence of switches/protective devices on a given upstream path; as well as a set of switches that satisfy a condition related to downstream components.

Chapter 7

Conclusions & Future Work

7.1 Discussion and Conclusion

The development of techniques for monitoring and data collection, and leveraging advanced data analysis methods provide the opportunity for thorough modeling and management of grid operations. It also creates a suitable circumstances for improving service quality in response to power outages, adverse weather conditions, and power equipment failures.

In Chapter 3, we investigated the power system protection challenges in the face of increasing penetration of wind farms into the power grid. We developed a sample 'test' system for objective comparison of different protection algorithms. In addition, we discussed various reasons and configurations that may lead to the malfunction of the protection system.

We analyzed several solutions proposed in the literature to overcome protection related challenges. We found out that understanding of the problems that the power system may experience has significant importance for the power utilities to prepare themselves while integrating the grid with renewable energy resources such as wind farms. If not addressed in advance, these challenges can affect the expected behavior of the protection system, as well as the location, severity, and impact of power outages. The main contributions of this chapter can be summarized as follows.

- Development of a sample 'test' system to discuss the main challenges of protection system in the power grids with integrated WFs.

- Investigation of the various configurations of fault locations and WFs to address the protection challenges.

- Categorization of the protection challenges and the proposed solutions for power systems with integrated WFs.

In Chapter 4, we provided a comprehensive overview of the power utility outage management system. We put much effort into collecting, understanding, cleaning, integrating, and processing all related data – the power system data, power outage data, and weather data. Furthermore, we provided new insights and statistics about power outages and their relationship with the power system and weather conditions.

We presented the Bayesian-based analysis of power outage in order to determine/update the power outage probability with the high spatial and temporal resolution based on the available weather data. The Bayesian approach provided valuable insight into the historical events that happened in the power grid.

Due to the availability of the weather data at the level of the 10x10 km grid, we defined a unique random variable to describe the posterior probability of power outage for each grid cell. This allowed us to determine probabilities that take into account the unique characteristics of the individual cells. We further demonstrated that the maximum posterior classifier model can be used as a classifier to distinguish between power outages and no power outages.

The main contributions of this chapter are:

- An overview of the utility outage management system's (OMS) database, through the data integration process, and demonstrated the relationship between various data sources.

- New insights and statistics on various types of power outages.

- Investigation of the interactions between weather, power system, and power outage data by transforming power outage data at the grid cells level and integrating it with weather data.

- Formulation of the predictive inferences based on posterior predictive distribution of the power outages and the weather condition.

In Chapter 5, we developed data-driven models to predict potential weather-related power outages. We constructed a reasoning framework that integrated pieces of evidence from various models under uncertainty using Dempster-Shafer's theory (DST).

Based on this, a system for predicting power outages and their type by combining multiple machine learning models had been developed for real-time outage prediction. A particular emphasis was placed on selecting the most suitable prediction model as well as on the overall architecture of the system. DST-based ensemble architecture provided the best results. Furthermore, this architecture allowed the user to look 'inside' the obtained results and gain a deeper understanding of how predictions were made. As a summary, the main contributions are as follows:

- Investigation of data-driven models to predict different weather-related power outages and their causes.

- Development of a reasoning framework under uncertainty using Dempster-Shafer theory (DST) for combining evidence from various sources is developed.

- Construction of a real-time weather outage prediction system by combining multiple models with a reasoning framework for predicting power outages.

The comparison of the probabilistic approach, Chapter 4, and machine learning models, Chapter 5, can provide a deeper insight into their differences in the context of their utilization.

The probabilistic model aimed to determine the probability of power outages based on the available data. We further discussed the cost function to assign a higher penalty for certain inaccurate predictions and to use the posterior probability as a classifier. However, in practice, it may not be applicable when there are not sufficient data points. Therefore, in Chapter 5, we introduced machine learning models as classifiers. These models use functional form to learn the probability distribution $p(o|\mathbf{w}, f)$. For these models, it is important challenge to make sure that they generalize well and have good performance on new unseen data. We tuned the hyper-parameters to minimize the generalization error to change the model capacity and prevent underfitting and overfitting.

The advantage of the power outage prediction system developed in Chapter 5 compared to Chapter 4 is its ability to generalize to unseen data. Since we tuned the parameters of the functional form of the probability distribution, the 'chapter 5 model' can interpolate or extrapolate based on the training data and provide inference on the unseen data.

The next important difference is how these models handle the unbalanced power outage data. In Chapter 4, we discussed the cost function as a means to put more penalties on specific inaccurate predictions. However, in machine learning models, although modifying the cost function is an applicable solution, under-sampling, over-sampling, and the utilization of performance metrics such as precision, recall, and F_1 are important techniques to deal with unbalanced data sets.

In Chapter 6, a power grid knowledge graph (GridKG) was developed. It had the capability to integrate grid topology information, data from customers, and information about power outages. In addition, algorithms were developed for identifying characteristics of the power grid and obtaining more insights from them. We enriched the GridKG with information such as the electrical paths in the grid, identifying the upstream and downstream paths, and pre-calculating the number of customers supplied through each specific device.

Furthermore, we demonstrated that GridKG is an effective tool for performing atypical analysis of grid behaviour. We used GridKG to perform Monte-Carlo simulations in order to estimate the impact of power outages and to determine distributions of affected customers in reference to their location. GridKG gives an easy way to access the topology of a power system and historical outage data at the same time – all of this allows to analyze impact of power outages on customers distributed across service areas while taking into account historically-based probabilities of equipment failures.

The main contributions of this chapter are as follows:

- Development of a power grid knowledge graph (GridKG) that integrates the grid topology data, with equipment, customers, and power outage data

- Design and implementation of algorithms to enhance the GridKG and generate more in-depth insight from the power grid characteristic

- Illustration of the GridKG ability in performing Monte-Carlo simulation for estimating the impact of power outage severity and providing probabilistic results

We can state that the proposed graph representation of the distribution system creates a unique environment for the processing and analysis of system data. It allows for a variety of algorithms to be deployed on it; it can be integrated and used with multiple data-driven techniques, such as a Monte-Carlo simulation.

7.2 Future Work

In general, this thesis represents a very important contribution to future activities related to power system outage analysis and utilizing data-driven methods and knowledge graphs in predicting weather-related power outages and estimating outage severity impacts. As a result of the proposed approaches in this research, several topics can be addressed and investigated in future studies.

7.2.1 Power System Protection Challenges

In the distribution sector, photovoltaic (PV) power generation has increased rapidly, resulting in more importance of PV system protection. The future work can focus on investigating the challenges imposed by PV energy penetration into the grid. Furthermore, it will be an interesting topic to compare the protection challenges and proposed solutions with systems with integrated WFs.

7.2.2 Outage and Weather Data Analysis

There are some potential topics for discussion in future studies to fully utilize the discussed approaches in Chapter 4. The posterior power outage probability estimation can be further developed to incorporate additional features such as vegetation status and equipment maintenance status. Furthermore, it can be implemented for various types of power outages. Different cost functions can be investigated, and the results can be compared with power outage estimation models based on the DST-based architecture. Moreover, The various time window sizes, weights, and window value functions can be examined further, and their impact on the posterior probability of power outages can be discussed.

7.2.3 Weather Outage Prediction system – *WoutPS*

Future work can focus on the DST-based fusion algorithm and investigate the local conflict and weight of conflict between pieces of evidence to estimate final confidence in *WoutPS* outputs. Furthermore, designing and developing a self-adapting subsystem for the continuous update of the prediction system can be another valuable topic for future studies.

7.2.4 Knowledge Graph for Power System

The concept of using knowledge graphs to represent grid data has numerous advantages compared to conventional data representation methods. Here are some suggestions for future research on this topic.

Integration of real-time data from power system measurement devices into the GridKG and commutation with protection relays can enhance the utilization of GridKG. It can be deployed into the outage management system and provide real-time severity and impact analysis of the ongoing power outages in the grid.

References

[1] R. J. Campbell and S. Lowry, "Weather-related power outages and electric system resiliency," Congressional Research Service, Library of Congress Washington, DC, 2012.

[2] D. Zhu, D. Cheng, R. P. Broadwater, and C. Scirbona, "Storm modeling for prediction of power distribution system outages," *Electric power systems research*, vol. 77, no. 8, pp. 973–979, 2007.

[3] R. Billinton and R. N. Allan, "Power-system reliability in perspective," *Electronics and Power*, vol. 30, no. 3, pp. 231–236, 1984.

[4] ——, "Basic power system reliability concepts," *Reliability Engineering & System Safety*, vol. 27, no. 3, pp. 365–384, 1990.

[5] E. National Academies of Sciences, Medicine, *et al.*, *Enhancing the resilience of the Nation's electricity system.* National Academies Press, 2017.

[6] A. Berkeley, M. Wallace, and C. COO, "A framework for establishing critical infrastructure resilience goals," *Final Report and Recommendations by the Council, National Infrastructure Advisory Council*, 2010.

[7] C. Office, *Keeping the country running: Natural hazards and infrastructure*, 2011.

[8] X. Yu and C. Singh, "Power system reliability analysis considering protection failures," in *IEEE Power Engineering Society Summer Meeting,*, IEEE, vol. 2, 2002, pp. 963–968.

[9] Y. Wang, C. Chen, J. Wang, and R. Baldick, "Research on resilience of power systems under natural disasters?a review," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 1604–1613, 2015.

[10] M. Panteli and P. Mancarella, "Modeling and evaluating the resilience of critical electrical power infrastructure to extreme weather events," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1733–1742, 2015.

[11] R. Eskandarpour and A. Khodaei, "Machine learning based power grid outage prediction in response to extreme events," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3315–3316, 2016.

[12] A. Pradhan and G. Joos, "Adaptive distance relay setting for lines connecting wind farms," *IEEE Transactions on Energy Conversion*, vol. 22, no. 1, pp. 206–213, 2007.

[13] S. Srivastava, U. Shenoy, A. Chandra Biswal, and G. Sethuraman, "Impedance seen by distance relays on lines fed from fixed speed wind turbines," *International Journal of Emerging Electric Power Systems*, vol. 14, no. 1, pp. 17–24, 2013.

[14] R. Dubey, S. R. Samantaray, and B. K. Panigrahi, "Adaptive distance protection scheme for shunt-facts compensated line connecting wind farm," *IET Generation, Transmission & Distribution*, vol. 10, no. 1, pp. 247–256, 2016.

[15] ——, "Simultaneous impact of unified power flow controller and off-shore wind penetration on distance relay characteristics," *IET Generation, Transmission & Distribution*, vol. 8, no. 11, pp. 1869–1880, 2014.

[16] R. Dubey, S. Samantaray, B. Panigrahi, and G. Venkoparao, "Adaptive distance relay setting for parallel transmission network connecting wind farms and upfc," *International Journal of Electrical Power & Energy Systems*, vol. 65, pp. 113–123, 2015.

[17] L. Tripathy, M. K. Jena, and S. Samantaray, "Differential relaying scheme for tapped transmission line connecting upfc and wind farm," *International Journal of Electrical Power & Energy Systems*, vol. 60, pp. 245–257, 2014.

[18] A. Hooshyar, M. A. Azzouz, and E. F. El-Saadany, "Distance protection of lines emanating from full-scale converter-interfaced renewable energy power plants-part i: Problem statement," *IEEE Transactions on Power Delivery*, vol. 30, no. 4, pp. 1770–1780, 2015.

[19] S. Xue, J. Yang, Y. Chen, C. Wang, Z. Shi, M. Cui, and B. Li, "The applicability of traditional protection methods to lines emanating from vsc-hvdc interconnectors and a novel protection principle," *Energies*, vol. 9, no. 6, p. 400, 2016.

[20] A. Hooshyar, M. A. Azzouz, and E. F. El-Saadany, "Distance protection of lines emanating from full-scale converter-interfaced renewable energy power plants—part ii: Solution description and evaluation," *IEEE Transactions on Power Delivery*, vol. 30, no. 4, pp. 1781–1791, 2015.

[21] G. M. Zubiri and S. L. Barba, "Impact on the power system protection of high penetration of wind farms technology," *CIGRE 2010*, vol. 5, no. 204, 2010.

[22] A. Hooshyar, M. A. Azzouz, and E. F. El-Saadany, "Distance protection of lines connected to induction generator-based wind farms during balanced faults," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 4, pp. 1193–1203, 2014.

[23] X. Chen, X. Yin, and Z. Zhang, “Impacts of dfig-based wind farm integration on its tie line distance protection and countermeasures,” *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 12, no. 4, pp. 553–564, 2017.

[24] S. De Rijcke, P. S. Pérez, and J. Driesen, “Impact of wind turbines equipped with doubly-fed induction generators on distance relaying,” in *Power and Energy Society General Meeting, 2010 IEEE*, IEEE, 2010, pp. 1–6.

[25] L. He, C.-C. Liu, A. Pitto, and D. Cirio, “Distance protection of ac grid with hvdc-connected offshore wind generators,” *IEEE Transactions on Power Delivery*, vol. 29, no. 2, pp. 493–501, 2014.

[26] D. Wanik, E. Anagnostou, B. Hartman, M. Frediani, and M. Astitha, “Storm outage modeling for an electric distribution network in northeastern usa,” *Natural Hazards*, vol. 79, no. 2, pp. 1359–1384, 2015.

[27] F. Yang, P. Watson, M. Koukoula, and E. N. Anagnostou, “Enhancing weather-related power outage prediction by event severity classification,” *IEEE Access*, vol. 8, pp. 60 029–60 042, 2020.

[28] D. Cerrai, D. W. Wanik, M. A. E. Bhuiyan, X. Zhang, J. Yang, M. E. Frediani, and E. N. Anagnostou, “Predicting storm outages through new representations of weather and vegetation,” *IEEE Access*, vol. 7, pp. 29 639–29 654, 2019.

[29] D. Wanik, J. Parent, E. Anagnostou, and B. Hartman, “Using vegetation management and lidar-derived tree height data to improve outage predictions for electric utilities,” *Electric Power Systems Research*, vol. 146, pp. 236–245, 2017.

[30] J. He, D. W. Wanik, B. M. Hartman, E. N. Anagnostou, M. Astitha, and M. E. Frediani, “Nonparametric tree-based predictive modeling of storm outages on an electric distribution network,” *Risk Analysis*, vol. 37, no. 3, pp. 441–458, 2017.

[31] D. B. McRoberts, S. M. Quiring, and S. D. Guikema, “Improving hurricane power outage prediction models through the inclusion of local environmental factors,” *Risk analysis*, vol. 38, no. 12, pp. 2722–2737, 2018.

[32] S. D. Guikema, R. Nateghi, S. M. Quiring, A. Staid, A. C. Reilly, and M. Gao, “Predicting hurricane power outages to support storm response planning,” *IEEE Access*, vol. 2, pp. 1364–1373, 2014.

[33] Y. Sun, A. K. Wong, and M. S. Kamel, “Classification of imbalanced data: A review,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 04, pp. 687–719, 2009.

[34] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, 2016, pp. 1310–1315.

[35] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[36] Y. Sasaki *et al.*, "The truth of the f-measure," *Teach Tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.

[37] N. Chinchor, "Muc-4 evaluation metrics," in *Proceedings of the 4th conference on Message understanding*, Association for Computational Linguistics, 1992, pp. 22–29.

[38] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems*, vol. 1, no. 1, pp. 3–28, 1978.

[39] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[40] G. Shafer, *A mathematical theory of evidence.* Princeton university press, 1976, vol. 42.

[41] J.-B. Yang and D.-L. Xu, "Evidential reasoning rule for evidence combination," *Artificial Intelligence*, vol. 205, pp. 1–29, 2013.

[42] M. A. Hady, F. Schwenker, and G. Palm, "Multi-view forest: A new ensemble method based on dempster-shafer evidence theory," *International Journal of Applied Mathematics and Statistics (IJAMAS): Special Issue on Soft Computing and Approximate Reasoning*, vol. 22, no. S11, pp. 2–19, 2011.

[43] T. Denoeux, "A neural network classifier based on dempster-shafer theory," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 30, no. 2, pp. 131–150, 2000.

[44] M. Lalmas, "Dempster-shafer's theory of evidence applied to structured documents: Modelling uncertainty," in *ACM SIGIR Forum*, ACM, vol. 31, 1997, pp. 110–118.

[45] P. Smets, "Decision making in the tbm: The necessity of the pignistic transformation," *International Journal of Approximate Reasoning*, vol. 38, no. 2, pp. 133–147, 2005.

[46] *Resource Description Framework*, `https://www.w3.org/RDF/`, Accessed: 2018-10-25.

[47] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic web*, vol. 8, no. 3, pp. 489–508, 2017.

[48] Q. Wang, Z. Mao, B. Wang, and L. Guo, “Knowledge graph embedding: A survey of approaches and applications,” *IEEE Trans. on Knowl. and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[49] H. Arnaout and S. Elbassuoni, “Effective searching of rdf knowledge graphs,” *Journal of Web Semantics*, vol. 48, pp. 66–84, 2018.

[50] R. Verborgh, M. Vander Sande, O. Hartig, J. Van Herwegen, L. De Vocht, B. De Meester, G. Haesendonck, and P. Colpaert, “Triple pattern fragments: A low-cost knowledge graph interface for the web,” *Journal of Web Semantics*, vol. 37, pp. 184–206, 2016.

[51] U. Lösch, S. Bloehdorn, and A. Rettinger, “Graph kernels for rdf data,” in *Extended Semantic Web Conference*, Springer, 2012, pp. 134–148.

[52] P. Szekely, C. A. Knoblock, J. Slepicka, A. Philpot, A. Singh, C. Yin, D. Kapoor, P. Natarajan, D. Marcu, K. Knight, *et al.*, “Building and using a knowledge graph to combat human trafficking,” in *International Semantic Web Conference*, Springer, 2015, pp. 205–221.

[53] T. Ruan, L. Xue, H. Wang, F. Hu, L. Zhao, and J. Ding, “Building and exploring an enterprise knowledge graph for investment analysis,” in *International Semantic Web Conference*, Springer, 2016, pp. 418–436.

[54] M. Needham and A. E. Hodler, *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O’Reilly Media, 2019.

[55] M. Tsili and S. Papathanassiou, “A review of grid code technical requirements for wind farms,” *IET Renewable power generation*, vol. 3, no. 3, pp. 308–332, 2009.

[56] G. Pannell, D. J. Atkinson, and B. Zahawi, “Minimum-threshold crowbar for a fault-ride-through grid-code-compliant dfig wind turbine,” *IEEE Transactions on Energy Conversion*, vol. 25, no. 3, pp. 750–759, 2010.

[57] O. Noureldeen, “Behavior of dfig wind turbines with crowbar protection under short circuit,” *International Journal of Electrical and Computer Sciences IJECS*, vol. 12, no. 3, pp. 32–37, 2012.

[58] S. Hu, X. Lin, Y. Kang, and X. Zou, “An improved low-voltage ride-through control strategy of doubly fed induction generator during grid faults,” *IEEE transactions on power electronics*, vol. 26, no. 12, pp. 3653–3665, 2011.

[59] J. J. Justo, F. Mwasilu, and J.-W. Jung, “Doubly-fed induction generator based wind turbines: A comprehensive review of fault ride-through strategies,” *Renewable and Sustainable Energy Reviews*, vol. 45, pp. 447–467, 2015.

[60] L. G. Meegahapola, T. Littler, and D. Flynn, “Decoupled-dfig fault ride-through strategy for enhanced stability performance during grid faults,” *IEEE Transactions on Sustainable Energy*, vol. 1, no. 3, pp. 152–162, 2010.

[61] K. E. Okedu, S. Muyeen, R. Takahashi, and J. Tamura, “Wind farms fault ride through using dfig with new protection scheme,” *IEEE Transactions on Sustainable Energy*, vol. 3, no. 2, pp. 242–254, 2012.

[62] J. John Justo and K.-S. Ro, “Control strategies of doubly fed induction generator-based wind turbine system with new rotor current protection topology,” *Journal of Renewable and Sustainable Energy*, vol. 4, no. 4, p. 043 123, 2012.

[63] J. Yang, J. E. Fletcher, and J. O’Reilly, “A series-dynamic-resistor-based converter protection scheme for doubly-fed induction generator during various fault conditions,” *IEEE Transactions on Energy Conversion*, vol. 25, no. 2, pp. 422–432, 2010.

[64] G. Pannell, B. Zahawi, D. J. Atkinson, and P. Missailidis, “Evaluation of the performance of a dc-link brake chopper as a dfig low-voltage fault-ride-through device,” *IEEE Transactions on Energy Conversion*, vol. 28, no. 3, pp. 535–542, 2013.

[65] M. Nasiri, J. Milimonfared, and S. Fathi, “A review of low-voltage ride-through enhancement methods for permanent magnet synchronous generator based wind turbines,” *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 399–415, 2015.

[66] J. Miret, A. Camacho, M. Castilla, L. G. de Vicuña, and J. Matas, “Control scheme with voltage support capability for distributed generation inverters under voltage sags,” *IEEE Transactions on Power Electronics*, vol. 28, no. 11, pp. 5252–5262, 2013.

[67] A. Camacho, M. Castilla, J. Miret, J. C. Vasquez, and E. Alarcón-Gallo, “Flexible voltage support control for three-phase distributed generation inverters under grid fault,” *IEEE transactions on industrial electronics*, vol. 60, no. 4, pp. 1429–1441, 2013.

[68] F. Elyasichamazkoti and S. Teimourzadeh, “Secure under frequency load shedding scheme with consideration of rate of change of frequency,” in *2021 IEEE Green Technologies Conference (GreenTech)*, IEEE, 2021, pp. 552–557.

[69] V. Cook, *Analysis of distance protection.* Research Studies Press, 1985, vol. 1.

[70] P. Dash, A. Pradhan, G. Panda, and A. Liew, “Adaptive relay setting for flexible ac transmission systems (facts),” *IEEE Transactions on Power Delivery*, vol. 15, no. 1, pp. 38–43, 2000.

[71] K. El-Arroudi, G. Joos, and D. T. McGillis, "Operation of impedance protection relays with the statcom," *IEEE Transactions on Power Delivery*, vol. 17, no. 2, pp. 381–387, 2002.

[72] T. S. Sidhu, R. K. Varma, P. K. Gangadharan, F. A. Albasri, and G. R. Ortiz, "Performance of distance relays on shunt-facts compensated transmission lines," *IEEE Transactions on Power delivery*, vol. 20, no. 3, pp. 1837–1845, 2005.

[73] X. Zhou, H. Wang, R. Aggarwal, and P. Beaumont, "Performance evaluation of a distance relay as applied to a transmission system with upfc," *IEEE Transactions on Power Delivery*, vol. 21, no. 3, pp. 1137–1147, 2006.

[74] F. A. Albasri, T. S. Sidhu, and R. K. Varma, "Performance comparison of distance protection schemes for shunt-facts compensated transmission lines," *IEEE Transactions on Power Delivery*, vol. 22, no. 4, pp. 2116–2125, 2007.

[75] M. Khederzadeh, "Upfc operating characteristics impact on transmission line distance protection," in *Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*, IEEE, 2008, pp. 1–6.

[76] K. Seethalekshmi, S. N. Singh, and S. C. Srivastava, "Synchrophasor assisted adaptive reach setting of distance relays in presence of upfc," *IEEE Systems Journal*, vol. 5, no. 3, pp. 396–405, 2011.

[77] M. Q. Duong, F. Grimaccia, S. Leva, M. Mussetta, G. Sava, and S. Costinas, "Performance analysis of grid-connected wind turbines," *The Journal Scientific Bulletin*, vol. 76, no. 4, pp. 169–180, 2014.

[78] K. El Arroudi and G. Joos, "Performance of interconnection protection based on distance relaying for wind power distributed generation," *IEEE Transactions on Power Delivery*, 2017.

[79] W. A. Qureshi and N.-K. C. Nair, "Wind farm protection," in *Large Scale Renewable Power Generation*, Springer, 2014, pp. 311–329.

[80] J. Morren and S. W. De Haan, "Short-circuit current of wind turbines with doubly fed induction generator," *IEEE Transactions on Energy conversion*, vol. 22, no. 1, pp. 174–180, 2007.

[81] T. K. Abdel-Galil, A. E. Abu-Elanien, E. El-Saadany, A. Girgis, Y. Mohamed, M. Salama, and H. Zeineldin, "Protection coordination planning with distributed generation," *Qualsys Engco. Inc*, 2007.

[82] V. A. Papaspiliotopoulos, G. N. Korres, V. A. Kleftakis, and N. D. Hatziargyriou, "Hardware-in-the-loop design and optimal setting of adaptive protection schemes for distribution systems with distributed generation," *IEEE Transactions on Power Delivery*, vol. 32, no. 1, pp. 393–400, 2017.

[83] R. C. Dugan and T. E. Mcdermott, “Distributed generation,” *IEEE Industry Applications Magazine*, vol. 8, no. 2, pp. 19–25, 2002.

[84] F. Coffele, C. Booth, A. Dyśko, and G. Burt, “Quantitative analysis of network protection blinding for systems incorporating distributed generation,” *IET Generation, Transmission & Distribution*, vol. 6, no. 12, pp. 1218–1224, 2012.

[85] B. Hussain, S. Sharkh, S. Hussain, and M. Abusara, “Integration of distributed generation into the grid: Protection challenges and solutions,” 2010.

[86] K. I. Jennett, C. D. Booth, F. Coffele, and A. J. Roscoe, “Investigation of the sympathetic tripping problem in power systems with large penetrations of distributed generation,” *IET Generation, Transmission & Distribution*, vol. 9, no. 4, pp. 379–385, 2014.

[87] L. Che, M. E. Khodayar, and M. Shahidehpour, “Adaptive protection system for microgrids: Protection practices of a functional microgrid system.,” *IEEE Electrification magazine*, vol. 2, no. 1, pp. 66–80, 2014.

[88] B. Ravindranath and M. Chander, *Power system protection and switchgear*. New Age International, 2011.

[89] A. Hooshyar, M. A. Azzouz, and E. F. El-Saadany, “Three-phase fault direction identification for distribution systems with dfig-based wind dg,” *IEEE Transactions on sustainable energy*, vol. 5, no. 3, pp. 747–756, 2014.

[90] M. Nagpal and C. Henville, “Impact of power electronic sources on transmission line ground fault protection,” *IEEE Transactions on Power Delivery*, 2017.

[91] J. Barsch, G. Bartok, G. BenmouyaI, O. Bolado, B. Boysen, S. Brahma, S. Brettschneider, Z. Bukhala, and J. Burnworth, “Fault current contributions from wind plants,” *A Report to the T&D Committee, Electric Machinery Committee and Power System Relaying Committee of the IEEE PES*, 2012.

[92] R. Walling, E. Gursoy, and B. English, “Current contributions from type 3 and type 4 wind turbine generators during faults,” in *Power and Energy Society General Meeting, 2011 IEEE*, IEEE, 2011, pp. 1–6.

[93] H. Sadeghi, “A novel method for adaptive distance protection of transmission line connected to wind farms,” *International Journal of Electrical Power & Energy Systems*, vol. 43, no. 1, pp. 1376–1382, 2012.

[94] C.-S. Chen, C.-T. Tsai, S.-C. Hsieh, C.-T. Hsu, and C.-H. Lin, “Adaptive relay setting for distribution systems considering operation scenarios of wind generators,” in *Industrial & Commercial Power Systems Technical Conf (I&CPS), 2013 IEEE/IAS 49th*, IEEE, 2013, pp. 1–8.

[95] Z. Liu, C. Su, H. Høidalen, and Z. Chen, "A multi-agent system based protection and control scheme for distribution system with distributed generation integration,"

[96] H. Yazdanpanahi, Y. W. Li, and W. Xu, "A new control strategy to mitigate the impact of inverter-based dgs on protection system," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1427–1436, 2012.

[97] H.-J. Lee, G. Son, and J.-W. Park, "Study on wind-turbine generator system sizing considering voltage regulation and overcurrent relay coordination," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 1283–1293, 2011.

[98] ——, "A study on wind-turbine generator system sizing considering overcurrent relay coordination with sfcl," *IEEE Transactions on Applied Superconductivity*, vol. 21, no. 3, pp. 2140–2143, 2011.

[99] A. Ghorbani, H. Mehrjerdi, and N. A. Al-Emadi, "Distance-differential protection of transmission lines connected to wind farms," *International Journal of Electrical Power & Energy Systems*, vol. 89, pp. 11–18, 2017.

[100] J. Mai, K. C. Kornelsen, B. A. Tolson, V. Fortin, N. Gasset, D. Bouhemhem, D. Schäfer, M. Leahy, F. Anctil, and P. Coulibaly, "The canadian surface prediction archive (caspar): A platform to enhance environmental modeling in canada and globally," *Bulletin of the American Meteorological Society*, vol. 101, no. 3, E341–E356, 2020.

[101] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning.* MIT press, 2016.

[102] W. H. Kersting, *Distribution system modeling and analysis.* CRC press, 2006.

[103] W. Kersting and R. Dugan, "Recommended practices for distribution system analysis," in *2006 IEEE PES Power Systems Conference and Exposition*, IEEE, 2006, pp. 499–504.

[104] H. C. Caswell, V. J. Forte, J. C. Fraser, A. Pahwa, T. Short, M. Thatcher, and V. G. Werner, "Weather normalization of reliability indices," *IEEE Transactions on Power Delivery*, vol. 26, no. 2, pp. 1273–1279, 2010.

[105] R. Eskandarpour and A. Khodaei, "Machine learning based power grid outage prediction in response to extreme events," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3315–3316, Jul. 2017, ISSN: 0885-8950. DOI: `10.1109/TPWRS.2016.2631895`.

[106] R. Eskandarpour and A. Khodaei, "Leveraging accuracy-uncertainty tradeoff in svm to achieve highly accurate outage predictions," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1139–1141, 2018.

[107] B. J. Cerruti and S. G. Decker, “A statistical forecast model of weather-related damage to a major electric utility,” *Journal of Applied Meteorology and Climatology*, vol. 51, no. 2, pp. 191–204, 2011.

[108] A. I. Sarwat, M. Amini, A. Domijan, A. Damnjanovic, and F. Kaleem, “Weather-based interruption prediction in the smart grid utilizing chronological data,” *Journal of Modern Power Systems and Clean Energy*, vol. 4, no. 2, pp. 308–315, 2016.

[109] S.-R. Han, S. D. Guikema, S. M. Quiring, K.-H. Lee, D. Rosowsky, and R. A. Davidson, “Estimating the spatial distribution of power outages during hurricanes in the gulf coast region,” *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 199–210, 2009.

[110] S.-R. Han, S. D. Guikema, and S. M. Quiring, “Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models,” *Risk Analysis: An International Journal*, vol. 29, no. 10, pp. 1443–1453, 2009.

[111] S. D. Guikema, S. M. Quiring, and S.-R. Han, “Prestorm estimation of hurricane damage to electric power distribution systems,” *Risk Analysis: An International Journal*, vol. 30, no. 12, pp. 1744–1752, 2010.

[112] Y. Zhou, A. Pahwa, and S.-S. Yang, “Modeling weather-related failures of overhead distribution lines,” *IEEE Transactions on power systems*, vol. 21, no. 4, pp. 1683–1690, 2006.

[113] H. Liu, R. A. Davidson, and T. V. Apanasovich, “Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms,” *Reliability Engineering & System Safety*, vol. 93, no. 6, pp. 897–912, 2008.

[114] K. Alvehag and L. Soder, “A reliability model for distribution systems incorporating seasonal variations in severe weather,” *IEEE Transactions on Power Delivery*, vol. 26, no. 2, pp. 910–919, 2010.

[115] P. Kankanala, A. Pahwa, and S. Das, “Estimation of overhead distribution system outages caused by wind and lightning using an artificial neural network,” in *Proc. Int. Conf. Power Syst. Oper. Plan.*, 2012.

[116] F. Xiao, J. D. McCalley, Y. Ou, J. Adams, and S. Myers, “Contingency probability estimation using weather and geographical data for on-line security assessment,” in *2006 International Conference on Probabilistic Methods Applied to Power Systems*, IEEE, 2006, pp. 1–7.

[117] Y. Liu and C. Singh, “A methodology for evaluation of hurricane impact on composite power system reliability,” *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 145–152, 2010.

[118] H. Liu, R. A. Davidson, and T. V. Apanasovich, “Statistical forecasting of electric power restoration times in hurricanes and ice storms,” *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 2270–2279, 2007.

[119] P. Gross, A. Boulanger, M. Arias, D. L. Waltz, P. M. Long, C. Lawson, R. Anderson, M. Koenig, M. Mastrocinque, W. Fairechio, *et al.*, "Predicting electricity distribution feeder failures using machine learning susceptibility analysis," in *AAAI*, 2006, pp. 1705–1711.

[120] C. Rudin, D. Waltz, R. N. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. N. Gross, B. Huang, S. Ierome, *et al.*, "Machine learning for the new york city power grid," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 2, pp. 328–345, 2011.

[121] P. Kankanala, A. Pahwa, and S. Das, "Mean field annealing based committee machines for outage estimation in power distribution system," in *Proc. 2nd Int. Association of Science and Technology for Development (IASTED) International Conf. Power and Energy Systems and Applications (PESA)*, 2012, pp. 12–14.

[122] P. Kankanala, S. Das, and A. Pahwa, "Adaboost$^+$: An ensemble learning approach for estimating weather-related outages in distribution systems," *IEEE Transactions on Power Systems*, vol. 29, no. 1, pp. 359–367, 2013.

[123] R. Nateghi, S. Guikema, and S. M. Quiring, "Power outage estimation for tropical cyclones: Improved accuracy with simpler models," *Risk analysis*, vol. 34, no. 6, pp. 1069–1078, 2014.

[124] S. D. Guikema and S. M. Quiring, "Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data," *Reliability Engineering & System Safety*, vol. 99, pp. 178–182, 2012.

[125] G. L. Tonn, S. D. Guikema, C. M. Ferreira, and S. M. Quiring, "Hurricane isaac: A longitudinal analysis of storm characteristics and power outage risk," *Risk analysis*, vol. 36, no. 10, pp. 1936–1947, 2016.

[126] S. M. Quiring, A. B. Schumacher, and S. D. Guikema, "Incorporating hurricane forecast uncertainty into a decision-support application for power outage modeling," *Bulletin of the American Meteorological Society*, vol. 95, no. 1, pp. 47–58, 2014.

[127] P.-C. Chen and M. Kezunovic, "Fuzzy logic approach to predictive risk analysis in distribution outage management," *IEEE Transactions on Smart Grid*, vol. 7, no. 6, pp. 2827–2836, 2016.

[128] A. Jaech, B. Zhang, M. Ostendorf, and D. S. Kirschen, "Real-time prediction of the duration of distribution system outages," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 773–781, 2018.

[129] C. Cortes, L. D. Jackel, and W.-P. Chiang, "Limits on learning machine accuracy imposed by data quality," in *Advances in Neural Information Processing Systems*, 1995, pp. 239–246.

[130] G. James and T. Hastie, “Generalizations of the bias/variance decomposition for prediction error,” *Dept. Statistics, Stanford Univ., Stanford, CA, Tech. Rep*, 1997.

[131] *Outages and weather*, `https://www.nbpower.com/en/outages/preparing-for-outages/outages-and-weather`, Accessed: 2019-10-11.

[132] N. Abi-Samra and W. Malcolm, “Extreme weather effects on power systems,” in *2011 IEEE Power and Energy Society General Meeting*, IEEE, 2011, pp. 1–5.

[133] *Kingston Utility report*, `https://utilitieskingston.com/News/Article/poles-collapse-report`, Accessed: 2018-07-25.

[134] M. Farzaneh, “Ice accretions on high–voltage conductors and insulators and related phenomena,” *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 358, no. 1776, pp. 2971–3005, 2000.

[135] M. Farzaneh, J. Zhang, and S. Aboutorabi, “Effects of insulator profile on the critical condition of ac arc propagation on ice-covered insulators,” in *Annual Report Conference on Electrical Insulation and Dielectric Phenomena*, IEEE, 2002, pp. 383–387.

[136] S. Ale-Emran and M. Farzaneh, “Flashover performance of ice-covered post insulators with booster sheds using experiments and partial arc modeling,” *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 23, no. 2, pp. 979–986, 2016.

[137] J. Hu, B. Yan, S. Zhou, and H. Zhang, “Numerical investigation on galloping of iced quad bundle conductors,” *IEEE Transactions on Power Delivery*, vol. 27, no. 2, pp. 784–792, 2012.

[138] M. Kermani, M. Farzaneh, and L. E. Kollar, “The effects of wind induced conductor motion on accreted atmospheric ice,” *IEEE Transactions on power delivery*, vol. 28, no. 2, pp. 540–548, 2013.

[139] C. Fan and X. Jiang, “Analysis of the icing accretion performance of conductors and its normalized characterization method of icing degree for various ice types in natural environments,” *Energies*, vol. 11, no. 10, p. 2678, 2018.

[140] M. Farzaneh and J. Kiernicki, “Flashover problems caused by ice build up on insulators,” *IEEE Electrical Insulation Magazine*, vol. 11, no. 2, pp. 5–17, 1995.

[141] K. Kannus and K. Lahti, “Laboratory investigations of the electrical performance of ice-covered insulators and a metal oxide surge arrester,” *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 14, no. 6, pp. 1357–1372, 2007.

[142] M. Farzaneh and O. Melo, "Properties and effect of freezing rain and winter fog on outline insulators," *Cold Regions Science and Technology*, vol. 19, no. 1, pp. 33–46, 1990.

[143] C. Zachariades, "Development of an insulating cross-arm for overhead lines," Ph.D. dissertation, The University of Manchester (United Kingdom), 2014.

[144] Y. Kor, M. Z. Reformat, and P. Musilek, "Predicting weather-related power outages in distribution grid," in *2020 IEEE Power & Energy Society General Meeting (PESGM)*, IEEE, 2020, pp. 1–5.

[145] *Cross-validation: Evaluating estimator performance*, `https://scikit-learn.org/stable/modules/cross_validation.html`, Accessed: 2019-10-11.

[146] Y. Tang, T. Liu, G. Liu, J. Li, R. Dai, and C. Yuan, "Enhancement of power equipment management using knowledge graph," in *2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)*, IEEE, 2019, pp. 905–910.

[147] R. Navigli and S. P. Ponzetto, "Babelnet: Building a very large multilingual semantic network," in *Proc. of the 48th annual meeting of the association for Comput. linguistics*, 2010, pp. 216–225.

[148] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*, Springer, 2007, pp. 722–735.

[149] *A lexical database for english*, `https://wordnet.princeton.edu/`, Accessed: 2020-08-11.

[150] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proc. of the 2012 ACM SIGMOD Int. Conf. on Management of Data*, 2012, pp. 481–492.

[151] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proc. of the 20th ACM SIGKDD Int. Conf. on Knowl. discovery and data mining*, 2014, pp. 601–610.

[152] Y. Kor, L. Tan, M. Z. Reformat, and P. Musilek, "Gridkg: Knowledge graph representation of distribution grid data," in *2020 IEEE Electric Power and Energy Conference (EPEC)*, IEEE, 2020, pp. 1–5.

[153] H. Huang, Z. Hong, H. Zhou, J. Wu, and N. Jin, "Knowledge graph construction and application of power grid equipment," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[154] T. Hubauer, S. Lamparter, P. Haase, and D. M. Herzig, "Use cases of the industrial knowledge graph at siemens.," in *Int. Semantic Web Conf. (P&D/Industry/BlueSky)*, 2018.

[155] B. Kan, W. Zhu, G. Liu, X. Chen, D. Shi, and W. Yu, “Topology modeling and analysis of a power grid network using a graph database,” *Int. J. of Comput. Intell. Systems*, vol. 10, no. 1, pp. 1355–1363, 2017.

[156] Z. Su, M. Hao, Q. Zhang, B. Chai, and T. Zhao, “Automatic knowledge graph construction based on relational data of power terminal equipment,” in *2020 5th Int. Conf. on Comput. and Communication Systems (ICCCS)*, IEEE, 2020, pp. 761–765.

[157] A. Devanand, G. Karmakar, N. Krdzavac, R. Rigo-Mariani, Y. F. Eddy, I. A. Karimi, and M. Kraft, “Ontopowsys: A power system ontology for cross domain interactions in an eco industrial park,” *Energy and AI*, vol. 1, p. 100 008, 2020.

[158] A. Perçuku, D. Minkovska, and L. Stoyanova, “Modeling and processing big data of power transmission grid substation using neo4j,” *Procedia Comput. science*, vol. 113, pp. 9–16, 2017.

[159] W. Zhiqiang, W. Yuan, Z. Kang, W. Xin, and H. Hui, “Entity alignment method for power data knowledge graph of semantic and structural information,” in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 569, 2019, p. 052 103.

[160] H. Huang, Y. Chen, B. Lou, Z. Hongzhou, J. Wu, and K. Yan, “Constructing knowledge graph from big data of smart grids,” in *2019 10th Int. Conf. on Information Technology in Medicine and Education (ITME)*, IEEE, 2019, pp. 637–641.

[161] Y. Yang, Z. Chen, J. Yan, Z. Xiong, J. Zhang, Y. Tu, and H. Yuan, “Multi-source heterogeneous information fusion of power assets based on knowledge graph,” in *2019 IEEE Int. Conf. on Service Operations and Logistics, and Informatics (SOLI)*, IEEE, 2019, pp. 213–218.

[162] B. Cui, “Electric device abnormal detection based on iot and knowledge graph,” in *2019 IEEE Int. Conf. on Energy Internet (ICEI)*, IEEE, 2019, pp. 217–220.

[163] S. Fan, X. Liu, Y. Chen, Z. Liao, Y. Zhao, H. Luo, and H. Fan, “How to construct a power knowledge graph with dispatching data?” *Scientific Programming*, vol. 2020, 2020.

[164] E. Kabir, S. D. Guikema, and S. Quiring, “Predicting thunderstorm-induced power outages to support utility restoration,” *IEEE Transactions on Power Systems*, 2019.

[165] F. Elyasichamazkoti, F. Aminifar, and M. Davarpanah, “Digital filter-based grid synchronization for autonomous microgrids,” *IET Renewable Power Generation*, 2021.

[166] Y. Kor, M. Davarpanah, R. Bekhradian, and M. Sanaye-Pasand, “Mitigating islanded mode small scale synchronous generator mechanical oscillations caused by electrical arc furnace,” in *2020 IEEE Electric Power and Energy Conference (EPEC)*, IEEE, 2020, pp. 1–8.