PREDICTIVE RISK ASSESSMENT FOR OPTIMAL ASSET MANAGEMENT IN POWER SYSTEMS

A Dissertation

by

TATJANA DOKIC

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Mladen Kezunovic
Committee Members,	Le Xie
	Serap Savari
	Jianhua Huang
Head of Department,	Miroslav Begovic

May 2019

Major Subject: Electrical Engineering

ABSTRACT

The electrical grid is among the most critical of infrastructures, as it both assures high quality of life and promotes economic growth. Loss of power leads to major economic, social, and environmental impacts. Annual impacts in the U.S. from weather-related outages in the electrical network result in more than $150 billion in lost revenue.

Due to the high level of environmental exposure of the electric utility overhead infrastructure, the most dominant cause of electricity outages is weather impact. About 75% of power outages are either directly caused by weather-inflicted faults (e.g., lightning), or indirectly due to weather-caused increases in equipment deterioration rates (e.g. insulation) leading to subsequent failures. While it is not feasible to prevent severe weather conditions, the impact of severe weather can be significantly reduced, and in some cases even eliminated, by accurate prediction of where faults may occur and what equipment may be vulnerable so that adequate maintenance or replacement mitigation approaches can be deployed.

The assessment of weather impact on electrical grids falls into a group of problems referred as "the Big Data problems". The electric power industry has been deploying a smart grid infrastructure containing anywhere from thousands to millions of measurement points throughout the network. In addition, comprehensive analysis of data not coming from utility infrastructure, such as weather, lightning, vegetation, and geographical base data, which also comes in great volumes, is necessary. This brings out new challenges in dealing with extremely large data sets and using them to improve decision-making. Efficient, predictive, condition-based asset management and outage mitigation requires real-time processing of large volumes of multi-domain data.

The goal of this research is to provide a comprehensive framework for the use of Big Data to assess weather impacts on utility assets. This is accomplished in four primary steps: 1) identifying the relevant weather parameters in relation to the electric power outages and asset deterioration; 2) evaluating electrical grid vulnerability to hazardous weather conditions using advanced data analytics; 3) predicting the risk imposed on the electrical grid by severe weather based on weather forecasts and network vulnerability estimates; and 4) developing optimal mitigation strategies that minimize the risk of weather-related power outages.

In this study a unified data framework is developed that enables collection and spatiotemporal correlation of variety of data sets. The Gaussian Conditional Random Fields (GCRF) algorithm is used for predicting the probability of future outages in the network, given weather forecast data. The temporal and spatial interdependencies between components and events in the network are leveraged for the improvement of prediction algorithm accuracy, and its capability to deal with bad and missing data. The algorithm shows high accuracy of prediction by predicting risk of 64% or higher for all the cases of outages in distribution, and over 74% for all cases in transmission. The prediction results are presented on a geographical map in the form of the Risk Maps. A dynamic asset management system based on optimization is built to reduce the predicted risk of outages and component failure while maintaining predetermined economic investment in periodic asset maintenance. The study approach is tested on real utility data for multiple applications. Two scenarios are observed in this dissertation to demonstrate the benefits of this approach: 1) lightning strikes on or in the vicinity of power lines; 2) the combination of high-speed wind and heavy precipitation causing lines to come in contact with surrounding vegetation.

DEDICATION

To my grandfather Emanuel Uvera, for teaching me how to repair, rebuild, and protect.

ACKNOWLEDGEMENTS

Thanks go to my family, to my mom Vesna, and my dad Stevo, for their love and encouragement. Special thanks to my brother Daniel and his wife Jelena. Without their help this would not be possible. Thanks to my love Tim, for all his help and support.

I wish to thank my advisor, Dr. Mladen Kezunovic, for his guidance and support throughout my studies at Texas A&M University. I would like to extend gratitude to my committee members, Dr. Le Xie, Dr. Serap Savari, Dr. Jianhua Huang, and Dr. James Caverlee, for their time, comments, and suggestions.

Thanks also go to my collaborators at Temple University, Prof. Zoran Obradovic, Dr. Jelena Gligorijevic, Dr. Djordje Gligorijevic, Shoumik Roychoudhury, and Martin Pavlovski, for their contributions in development of the predictive algorithms used for this research.

I would like to thank my friends and colleagues and the department faculty and stuff for providing great support during my time at Texas A&M University. I would especially thank my colleagues Po-Chen Chen and Qin Yan for their friendship and great teamwork.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Mladen Kezunovic [Advisor], Dr. Serap Savari and Dr. Le Xie of the Department of Electrical and Computer Engineering, Dr. Jianhua Huang of the Department of Statistics, and Dr. James Caverlee of the Department of Computer Science and Engineering.

The predictive algorithms used in this study were selected and developed with guidance from Prof. Zoran Obradovic and his research group at Temple University.

Funding Sources

My research was mainly founded by three sources. a) PSerc consortium, under the project "Systematic Integration of Large Data Sets for Improved Decision-Making"; b) CenterPoint Energy under the project "Vegetation Management Risk Model"; and c) NSF Grant "Smart Grid Big Data SPOKE." Additional support came from following scholarships and fellowships:

- Willemain, Guyonnot, Liagre '99 Scholarship, Texas A&M, 2017 – 2019;
- Thomas W. Powell '62 and Powell Industries Inc. Fellowship, Texas A&M, 2017/18;
- Hawaii International Conference on System Sciences Doctoral Fellowship, 2017;
- ISS International Education Fee Scholarship (IEFS), Texas A&M, 2017/18;
- Fund for Young Talents Dositeja, Ministry of Youth and Sports, Serbia, 2012 – 2018.

The dissertation's contents are solely the responsibility of the author and do not necessarily represent the official views of the above awarding offices.

NOMENCLATURE

ASOS	Automated Surface Observing System
BD	Big Data
BIL	Basic Lightning Impulse Insulation Level
CBM	Condition-Based Maintenance
DFR	Digital Fault Recorder
GCRF	Gaussian Conditional Random Fields
GIS	Geographic Information System
GPS	Global Positioning System
ICM	Intelligent Condition Monitor (includes Intelligent Transformer Monitor – ITM, Circuit Breaker Condition Monitor – BCM, etc.)
IED	Intelligent Electronic Device
LIDAR	Light Detection and Ranging
NAM	North American Mesoscale
NASA	National Aeronautics and Space Administration
NDFD	National Digital Forecast Database
NERC	North American Electric Reliability Corporation
NLDN	National Lightning Detection Network
NOAA	National Oceanic and Atmospheric Administration
NTP	Network Time Protocol
NWP	Numerical Weather Prediction
OT	Optimization Technique

PMU	Phasor Measurement Unit
PTP	Precision Time Protocol
Radar	Radio Detection and Ranging
RCM	Reliability-Centered Maintenance
RTF	Run-To-Failure
Satellite	Geostationary and Polar-Orbiting Meteorological Spacecraft
SM	Smart Meter
TAI	International Atomic Time
TNRIS	Texas Natural Resources Information System
TPWD EMST	Texas Parks & Wildlife Department - Ecological Mapping Systems of Texas
UTC	Coordinated Universal Time
UV	Ultraviolet
WRF	Weather Research and Forecast
WFM	Weather Forecast Model

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION: PROBLEM STATEMENT*

1.1 Introduction

Power system outages have significant impacts on both the economy and the quality of human life. Large blackouts can cause major economic losses, as well as losses in human lives and severe injuries. Compared to large blackouts, smaller outages can cause more cumulative economic losses over periods of several years, due to their high frequency of occurrence. Results of various outage cause surveys conducted in recent years are presented in Fig. 1 and Fig. 2, [1, 2]. Based on these studies we can conclude that the majority of power outages (55% in Fig. 1 and 76% in Fig. 2) are caused by a) severe weather, which can lead to trees touching the conductors, or lightning and other environmental hazards causing a temporary loss of insulation properties, and b) permanent equipment failure caused by deterioration and eventual breakdown of the equipment electrical insulation strength. Due to the high level of exposure of the network overhead equipment to environmental impacts, equipment failure rates are highly correlated with weather impacts [3].

1.2 Weather Impacts

If we observe the cause of the power system outages, we can identify two main types of environmental impacts [4]:

1. Instantaneous impact to utility assets: All outages that are directly caused by environmental impact are placed in this group. Some examples of this type of outage

* This section is in part a reprint with permission of the material in the following paper: M. Kezunovic, T. Dokic, “Predictive Asset Management Under Weather Impacts Using Big Data, Spatiotemporal Data Analytics and Risk Based Decision-Making,” 10th Bulk Power Systems Dynamics and Control Symposium – IREP’2017, Espinho, Portugal, August 2017.

are lightning strikes causing an insulator backflashover, severe storms breaking or moving the conductors against each other, and tree branches growing into the lines and causing short circuit. When post-fault analysis is performed, such types of outages are

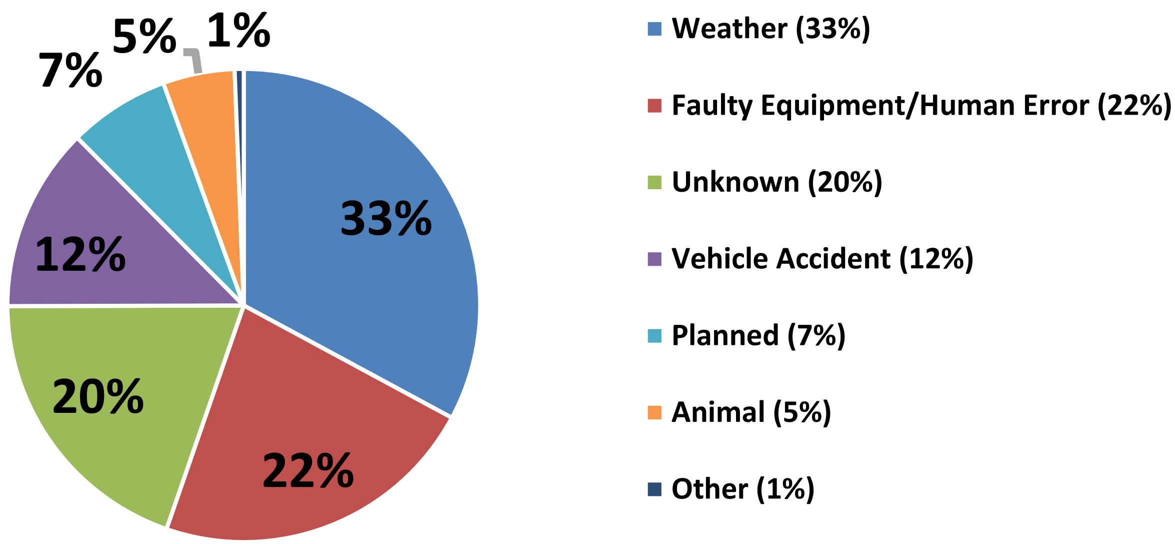

Figure 1 Power Outages by Cause in the USA in 2017 by Eaton, adapted from [1]

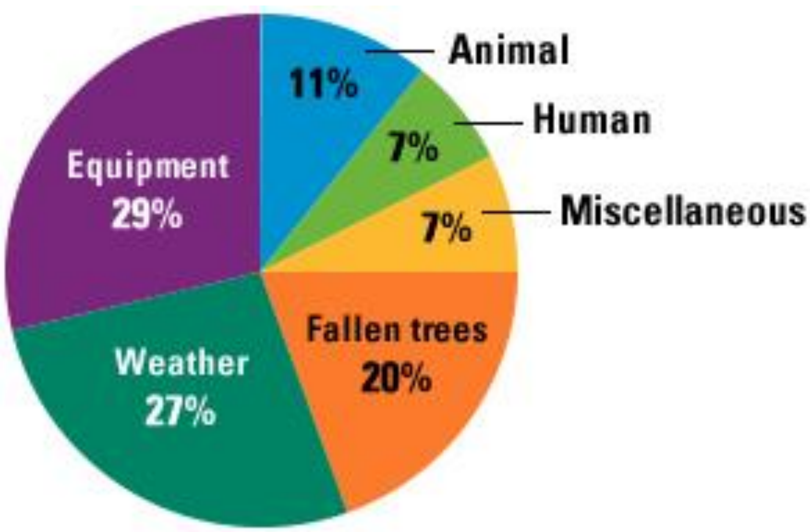

Figure 2 Power Outages by Cause by We Energies, reprinted from [2]

marked as being caused by a specific type of weather event such as lightning, ice, wind, or in general as weather induced.

2. Cumulative impact on utility assets: In this group we include outages that occurred due to equipment insulation deterioration caused by prolonged exposure of utility assets to weather impacts. Sustained weather impact on equipment can accelerate component deterioration. Some of the examples are an increase in demand caused by hot weather conditions that lead to overloading of utility equipment, or an exposure to multiple lightning strikes over time that cause the performance of lightning protection equipment (insulators, surge arresters) to deteriorate quicker. During the post-fault analysis such types of outages are typically marked as equipment failure, and sometimes it is not possible to directly correlate the outage with its cause. However, with statistical analysis it is possible to analyze how exposure to different weather conditions over time impacts the component's reliability.

We use three terms to describe different levels of weather impact: severe, catastrophic, and extreme. **Severe weather** impact describes situations where there is a chance that the weather conditions will cause some damage to the infrastructure, or endanger human life [5]. Examples of severe weather are thunderstorms, tornadoes, hail, storms, etc. A special case of severe weather is **catastrophic weather**, characterized by the extremely high potential for causing damage, serious social disruptions, and potential loss of human life [6]. Examples of catastrophic weather are tornadoes, hurricanes, earthquakes, tsunamis. **Extreme weather** describes the event where one or multiple weather parameters has reached an extreme value compared to its historical distribution. Examples are extreme wind speed and/or gust, extreme temperature, extreme cold waves, etc. [7].

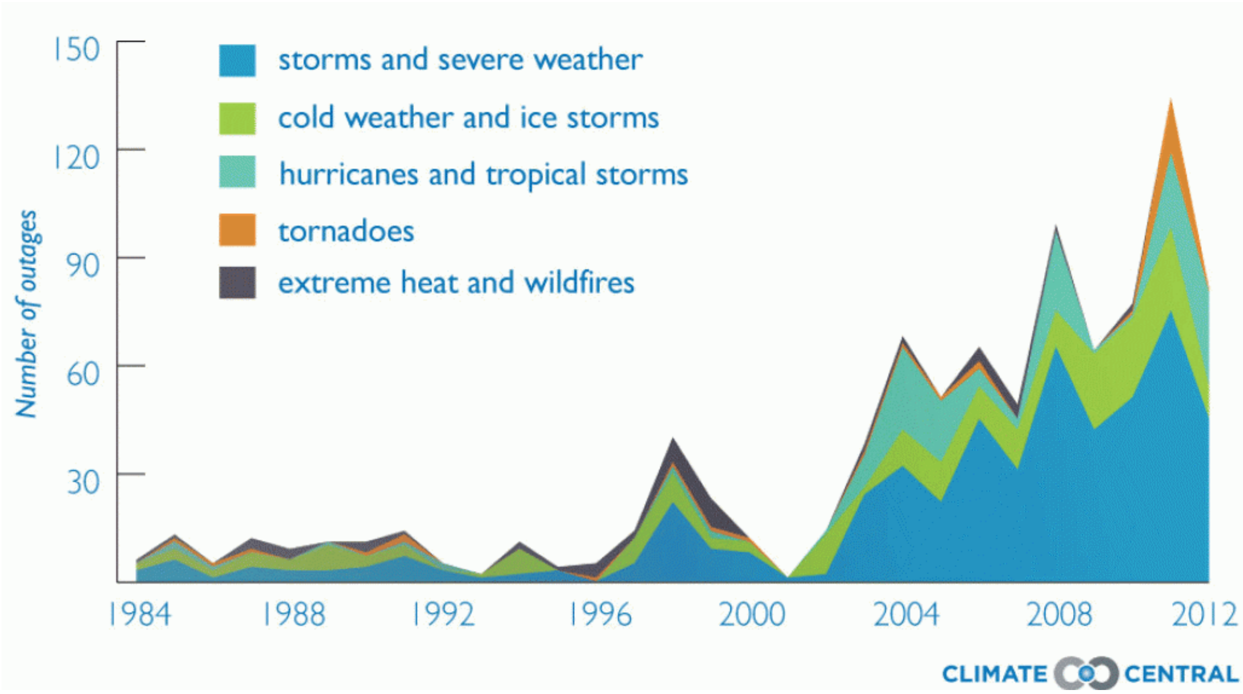

Figure 3 Trends of Weather-Related Power Outages by Climate Central for major events including at least 50,000 customers affected, reprinted from [9]

This research emphasizes the prediction of severe weather impacts on the transmission and distribution system, since these events cause the largest overall economic impact on utility operation and management. Catastrophic weather accounts for only 7% of large blackouts [8], while more than 50% are due to frequent storms and severe weather. As presented in Fig. 3 [9], the number of power outages due to weather impacts has been dramatically increasing in recent years, and storms and severe weather impacts are characterized by the fastest increase. This increase in severe weather events is resulting in huge economic, social, and environmental risks to power systems and its customers [10].

The US economy loses $104-$164 billion a year to outages and another $15-$24 billion to power quality impacts associated with loss-of-power surges [11]. The Northeast Blackout that occurred in August 2003 was caused by multiple trees making contact with transmission lines resulting in an erroneous line tripping [12]. The associated economic losses are estimated to be between $4 and $10 billion [13]. Another example of a large blackout occurred on December 2008 in New Hampshire, which started with ice-damaged tree falling on the lines during the ice storm. More than 800,000 people were affected [14].

A common weather impact that contributes to faults is the combination of high wind activity and trees being blown into the lines [15, 16]. Thus, it is of great interest for utilities to minimize the risk of vegetation-related outages in power system [17]. Tree trimming is the largest expense associated with distribution system management. Utilities spend millions of dollars on tree trimming every year [18].

Lightning is yet another common cause of outages [19]. Lightning strikes generate overvoltages that travel along the transmission lines affecting insulators. While they will not always cause the failure of equipment, their intensity and frequency of occurrence will affect the rate of insulation deterioration [20]. Evaluating the impact of lightning-caused overvoltages on the insulators along the transmission lines is of utmost importance. The insulators play an important role in power transmission system as the integrity of an overhead transmission line is directly governed by the electrical and mechanical performance of such equipment. While the insulators account for only 5% to 8% of the direct capital cost of the transmission line, more than 70% of the line outages and up to 50% of line maintenance costs are being caused by the insulator-induced outages [21].

1.3 Big Data

The measurements from the physical network and weather data exhibit various aspects of Big Data (BD). For efficient condition-based asset and outage management, and preventive real-time operation, fast processing and spatiotemporal correlation of large volumes of data is required. Table 1 lists characteristics of several BD sources that are of interest. The volume and velocity with which the data is generated can be overwhelming for both on-request and real-time applications. The heterogeneity of data sources and accuracy creates additional challenges.

Asset and outage management have relied heavily on model-based solutions in the past [22-27]. The inability of such methods to reflect the dynamically changing weather impacts over time makes it difficult to assess the deterioration of power grid infrastructure, anticipate locate faults, and predict operating conditions [19]. The advancements in smart grid measurement technologies have enabled the necessary conditions for the development of new data-driven solutions. While the new sensors can provide abundant information about the power system state, there is a lack of data analytics to extract knowledge needed to draw causal inferences between evolving power system events and the weather elements causing emergencies. The sufficient condition to derive such knowledge is to use adequate model-based approaches that directly incorporate Big Data. Thus, it is necessary to build a hybrid system that utilizes both physical model- and data-driven solutions as two complementary parts required to observe the evolving weather impacts on the power system. Data helps calibrate the models and defines necessary causal relations, while models help reduce the impact of missing or low-quality data.

1.4 Conclusion

We can conclude that assessing weather impacts on power systems and reducing the number of weather-related outages can lead to a substantial improvement of power system performance, from both economic and reliability perspectives. Recently, both utilities and outside agancies are collecting an excessive amount of data that can be used to analyze various aspects of weather relate outages. With the development of smart automatic data analytic tools, great value

Table 1 Challenges of Big Data for a typical utility

	Data Class	Data Source (Measurements)	VOLUME (Data file size)	VELOCITY (Rate of use)	VERACITY (Accuracy)
VARIETY	Utility measurements	SM	120GB per day/ device	Every 5-15 min	error <2.5%
		PMU	30GB per day/device	240 samples/sec	error <1%
		ICM	5GB per day/device	250 samples/sec	error <1%
		DFR	10MB per fault/device	1600 samples/sec	error <0.2%
	Weather data	Radar [28]	612 MB/day per radar	Every 4-10 min	1-2 dB; m s^{-1}
		Satellite [29]	At least 10 GB per day	Every 1-15 min	VIS<2%; IR<1-2K
		ASOS [30]	10 MB/day per station	Every 1 min	T-1.8°F, P<1%, Wind speed - 5%, RR - 4%
		NLDN [31]	40 MB/day	During lightning	SE < 200m, PCE <15%
		NDFD [32]	5-10 GB/day per model	1 - 12 hours	Varies by parameter
	Vegetation and Topography	TPWD EMST [33]	2.7 GB for Texas	static	SE < 10 m
		TNRIS [34]	300 GB for Texas	static	SE < 1 m
		LIDAR [35]	7 GB for Harris Co.	static	HE < 1m, VE < 150 cm

can be extracted from these data sources, allowing outage prediction and mitigation. However, there are many challenges that need to be addressed when using Big Data and associated data analytics for prediction and assessment of weather related outages.

Chapter II will address some of the existing solutions related to various aspects of weather related outages and Big Data applications in power systems. Then, beginning with Chapter III and continuing into the subsequent chapters, we will propose a data analytics framework that can be used to assess weather-related events in power systems for many different applications.

CHAPTER II

CURRENT PRACTICES AND LITERATURE REVIEW*

2.1 Introduction

In this chapter, we survey various literature sources to assess the existing analytics capabilities for using Big Data to predict outages in power systems. We need to analyze the solutions in different fields of study. First, we want to analyze the current state of the art in the area of asset management, which will provide a background for us to compare our methodology in Chapter IV. Then, to understand impacts of different types of outages, we are separating solutions related to two main causes of outages, vegetation and lightning, as well as general weather impact studies.

2.2 Asset Management

The traditional approach to asset condition monitoring is to perform laboratory tests to assess initial properties of the asset and its performance breakdown point [19]. Subsequently, the test results are used as a reference for periodic field assessments [36]. The frequency of field examination varies based on the device type and operating conditions. A different approach to assets management is "run-to-failure", where the components are never being inspected and actions are taken only after the component breaks [37]. In recent decades, technological advances have made it possible to closely monitor an asset's states and characteristics using periodic or continuous measurements from various sensors [38]. Today, these sensors are typically integrated

* This section is in part a reprint with permission of the material in the following papers: (1) M. Kezunovic, T. Dokic, "Predictive Asset Management Under Weather Impacts Using Big Data, Spatiotemporal Data Analytics and Risk Based Decision-Making," 10th Bulk Power Systems Dynamics and Control Symposium – IREP'2017, Espinho, Portugal, August 2017. (2) T. Dokic, M. Kezunovic, "Predictive Risk Management for Dynamic Tree Trimming Scheduling for Distribution Networks," IEEE Transactions on Smart Grid, September 2018. Copyright 2018, IEEE.

with intelligent electronic devices (IEDs) and provide a continuous on-line condition-based monitoring of equipment [39].

Another approach is a risk-based maintenance scheduling, illustrated by several recent studies [19]. In [40] the risk-based allocation of maintenance resources to various distribution system assets is described. The method uses linear optimization to balance risk reduction and economic losses. Research in [41] uses decoupled risk factors and a mixed-integer linear formulation for optimization of maintenance tasks. Work in [42, 43] demonstrates the application of a risk assessment analysis of a structured asset model with a function-oriented business process model. In [44] a nonparametric regression method is used to develop a failure rate model based on proportional hazards modeling. The study in [45] develops a risk assessment framework for extreme events caused by simultaneous or cascading faults.

We identify the following shortcomings in the mentioned works, which this dissertation has addressed:

- There is a lack of capability to process, utilize, and visualize larger amounts of diverse data in real-time. While there are a number of solutions for offline analysis of asset states [36,37], and asset diagnostics [38,39], there is a limited capability to process and utilize the large amount of data for asset condition prediction.
- Implementation of predictive analytics using spatiotemporal data is not considered. In most studies, the assets were analyzed as individual components, or the impacts were averaged over multiple devices of the same kind [40-43]. However, none of these studies take spatial configuration of assets as a factor for predicting the future state of the asset.

- Dynamic maintenance scheduling based on real-time observations of network components' states and the surrounding conditions is not enabled. Some studies do implement optimization based on risk reduction [40,41]. However, the risk in these studies was calculated statistically, where various impacts were averaged over multiple components, without taking into account the variety of measurements.

2.3 Weather Impacts

There have been some efforts to develop a weather impact assessment in recent years as reported in [19]. Time-varying weight factors were introduced as a measure of weather impact to component failure rates and restoration times [46]. Historical weather data were correlated with historical outage data to develop a damage forecast model for restoration [47]. A variety of studies have addressed the impact of extreme [48-50] and catastrophic [51, 52] weather on power system infrastructure. The impacts of large-scale storms and hurricanes have been evaluated [48], while risk analysis has been performed for evaluating wind storm impacts [49]. The impacts of Hurricane Sandy have been evaluated as suggested in [50].

A probabilistic framework for assessment of extreme weather conditions' impact on the grid [50], and system restoration after the extreme weather events [52] have been studied. There are two limitations of the existing weather impact methods that this research would like to address:

- Although existing solutions have good performance for improving the post-outage restoration process, the predictive capabilities that would enable proactive preventive maintenance and operation are missing.
- Most of the studies are focused on extreme and catastrophic events, while there is a lack of assessment for the weather impact of daily severe weather conditions.

2.4 Vegetation Management

For the prediction of vegetation potential to cause faults due to inadequate tree trimming practices, the most important factor is plant growth rate [53]. There are two types of models for estimating plant growth dynamics [54]: 1) process-based models that aim at defining the processes that cause the tree growth [55], and 2) data-based models that are empirical [56]. Distribution lines do not have strictly enforced right-of-way and often end up being placed in a close vicinity of the growing vegetation. Due to the high expenses of trimming large areas populated by many distribution feeders and neighborhood concerns for preserving nature, it is not economical or environmentally feasible to have all trees securely trimmed at all times.

In most cases, this process of tree trimming is applied in utilities based on a predetermined periodic schedule [57]. The only occasion when the schedule is changed is as a reactive measure to a vegetation-caused outage. The current practice relies on visual inspection made by crews using helicopters, airplanes, ground vehicles, or sometimes even people walking up to the lines [58]. Because of the high cost of this practice, it is of economic benefit to develop visual inspection methods that can provide automatic identification of dangerous zones [53].

Work in [59] uses Markov models to find the optimal inspection frequency while finding a compromise between the reliability of the system and the cost of distribution feeder inspection. Then, in [57], an optimal tree-trimming schedule is developed based on a hybrid genetic algorithm consisting of simulated annealing, genetic algorithms, and Tabu search. Vegetation-related failure rates are predicted using four different algorithms in [54]: linear regression, exponential regression, linear multivariable regression, and an artificial neural network. The developed predictors use historical outage data and only some of the weather parameters, neglecting the vegetation indices. In the listed literature, when weather impacts are considered, only few a variables of interest are included and their impact is averaged over time.

Two models, negative binomial generalized linear model and a Poisson generalized linear mixed model, were used in [60] to evaluate the impact of tree trimming to the rates of vegetation-caused outages in distribution. The data used in this study was limited to the utility-collected data, without insight into weather and vegetation indices. In [61, 62] satellite imagery was used to identify trees around the transmission lines. While the use of high-resolution imagery did show the potential in transmission vegetation management, its use in distribution networks was not discussed. Work in [63, 64] developed a reliability-centered vegetation management while looking closely into the electrical characteristics of vegetation-related outages. The work in [65] demonstrated the potential of spatial correlation of Big Data for the improvement in distribution vegetation management, but did not provide the necessary data analytics.

An analysis of the above methods identifies a number of shortcomings and constraints of the current data analytics approaches that we will try to overcome:

- The amount and variety of different data is very limited. Specifically, some studies do not use any vegetation indices, and are limited to utility-collected data and some limited weather data sources [54, 57, 59].
- There is no automatic way to observe the changes of the vegetation around the lines, and manual inspections are very expensive. While there were some efforts in creating automatic system for observation of vegetation progress around the transmission lines [61,62], no such studies were reported for distribution network. This is due to the higher spatial density of distribution compared to transmission, which makes it difficult to collect and process a variety of data sources with the appropriate resolution of data.

- Limited predictive capabilities are reported for vegetation-caused outages. The algorithms used in the studies were trained using limited input data [54, 57], and were not developed to support or exploit multiple diverse data sources, or the correlation of these data sources.
- There are no dynamic optimal tree trimming approaches based on predictive methods. There are some tree trimming or tree inspection studies [59, 60], but they do not use predictive models to estimate the state of vegetation around the lines.

2.5 Insulation Coordination

As reported in [19], lightning studies and experiences with insulation coordination have been reported in [66-70]. For the purpose of estimating the probability of a lightning strike, historical lightning data has been used in [66, 67]. Correlation of lightning data with transient measurements has been studied in literature [71, 72]. In [72], real time monitoring of transmission line transients during lightning strikes was presented, which allowed for spatiotemporal correlation of lightning data and transient measurements to evaluate the impact on insulation coordination.

In [70], lightning data is correlated with traveling wave fault locator data to provide better accuracy and robustness of a fault location algorithm. Correction factors for utilization of weather station data for insulation coordination have been described in [73]. In [74], a statistical method for lightning-related risk analysis was performed. An optimization procedure to determine locations of line arresters that minimize the risk was implemented. In [74], weather conditions were taken into account; however, the study has been performed based on randomly generated data. All of the methods were minimizing statistically-calculated risk functions considering insulator strength as defined by the insulator manufacturer.

What is missing in all of the studies is:

- There is no capability to observe the history of each individual insulator and make predictions based on that. Rather, studies do statistical analysis that averages the expectations over a large number of components [74], without enabling the prediction of an individual insulator's state based on its history.
- Weather impacts on the components are not tracked and recorded in real time. This means that in all of the studies the assumed level of insulator strength was wrong. The only study that is an exception is [73]; however, this study does not make predictions of insulators' future states based on weather forecast.
- Estimates of insulator strength are not updated with time based on the experienced lightning events and atmospheric conditions around the network. None of the studies report the capability to track the weather impacts on insulators in real-time, analyze them, and make predictions based on them.

2.6 Predictive Analysis for Power System Studies

Asset and outage management, as well as operations planning, have relied heavily on physical model-based solutions in the past [75-80]. The inability of such methods to reflect the dynamically changing impacts over time makes it difficult to assess the unfolding deterioration of power grid infrastructure, anticipate fault location, and predict operating conditions [19]. The advancements in smart grid measurement technologies have enabled the necessary conditions for development of new data-driven solutions. While new sensors can provide abundant information about the power system state, there is a lack of data analytics to extract knowledge needed to draw causal inferences between evolving power system events and emergencies caused by weather.

It is important to look into advancements in data analytics and identify the ways they can improve the existing power system applications and open the door for possible new solutions and capabilities. Such new technologies can help improve the reliability of the system with advanced prediction methods. These prediction methods can mitigate outages, improve the resilience of the system, and reduce restoration time and cost. The power system community can benefit from these approaches in the following ways:

- With more information coming from the new measurements, being collected in many domains surrounding network-related events, the accuracy of algorithms used for power system applications can be improved. In past studies, many of the impacts on the network were neglected due to the lack of data. For example, the use of weather conditions data during lightning-caused outages can improve the level of knowledge about the event that creates better input data for a prediction model.
- The predictive capabilities of these algorithms can be used to move the practice from the mostly reactive decision-making that is dominant today, to more predictive and proactive decision-making. If we are able to predict and prevent more outages, we can significantly improve the overall reliability of the system. In addition, smart decisions based on expected risk to outages around the network can result in a more strategic allocation and distribution of maintenance resources around the network. This, in turn, will result in faster restoration and improve the overall resilience of the system.
- Electric networks have experienced a number of changes in recent years, including the addition of multiple types of renewable energy sources such as concentrated and distributed solar and wind generators, and electric vehicle integration. Data analysis can provide an automatic platform to support the dynamics of these changes in the network. The renewable energy sources are highly dependent on environmental

impacts and conditions. Thus, analytics can greatly benefit from exploring the available data sources, and finding ways to analyze them in an efficient and accurate way.

In recent years, a variety of power system data analytics studies in the literature have incorporated data-driven approaches based on various data mining techniques: regression models [81-83], clustering and classification [84-86], support vector machines [87-88], neural networks [89-90], deep learning [91-92], etc. Regression models have shown great performance in various applications that analyze historical measurements to predict future events in the network through ether logistic or linear regression. Clustering and classification methods have found their place in event classification applications based on PMU data. Support Vector Machines have proven to be powerful in dynamic stability analysis based on synchrophasor data. Neural network solutions have been used in various applications, e.g., optimal maintenance scheduling and optimal placement of various components in the network. Deep learning techniques are finding their way into various applications for real-time load forecasting and emergency management.

The type of problem that is being solved in this dissertation is a regression problem, since we are trying to determine the relationship between a dependent variable (insulator strength, vegetation status, outage probability) and a number of input variables that are collected through various measurements (weather measurements, vegetation parameters, etc.). The choice of regression algorithm is based on consideration of the following three factors:

1) The collected inputs are not only observed in time, but they are also spatially distributed. This means that the solution could benefit a great deal from a structured regression predictor that is able to incorporate the spatial distribution. Due to the electric network being easily represented in a form of a connected graph, this is a

natural data structure to be chosen as the network representation. This results in the recommendation to use a structured regression algorithm.

2) The main reason driving the selection of prediction algorithm for this study comes from the fact that not only are the inputs observed spatially, but also there is a strong spatial correlation between outputs. In all of the studies considered in this research, a high probability of a given event in one node means that there is an increased probability of the same type of the event in neighboring nodes. For example, if there was a lightning strike on one tower in the network, the impact of the lightning wave propagates and can significantly affect all of the nearby nodes. It will, however, attenuate after a certain distance. Also, based on our historical outage data we were able to observe that in a majority of instances multiple weather-related outages will be clustered in a small area of the network. This results in the recommendation to use an algorithm that is capable of modeling the spatial interdependencies between outputs, which Conditional Random Fields can provide.

3) We are modeling each tower or each span as one node, which results in hundreds of thousands of nodes, with over 30 different parameters measured for thousands of events over the history of each component. The high computational complexity when applying structured regression on large power system networks can be avoided by constructing CRF feature functions as quadratic functions of outputs. This results in a final recommendation to use the Gaussian Conditional Random Fields algorithm that can satisfy all the specifics of the problem at hand: network structure, scalability of prediction, spatiotemporal relations between inputs and outputs, and computational efficiency.

2.7 Conclusion

While all the surveyed literature presents a variety of advanced methods for analysis of weather impacts on power systems outage and asset management, there are several aspects that are not addressed, and that will be taken into account in this dissertation.

- First, the variety of data used in these past studies was very limited. Even though a number of data sets became available in recent years, they did not find their way into power system studies due to the high complexity of collecting and integrating such data with traditional power system databases. We place an empahasis in this study on collecting a large amount of data coming from sources that have never been correlated in the past. This research describes the development process for combining such a diverse set of data and extracting knowledge from it in real-time.
- Second, current applications are mostly oriented towards statistical analysis of very limited data sets. This often results in averaging the impacts over a large number of components, and does not take into account specific events that occurred on individual devices over time. The main barrier to providing this capability is the development of a framework that can take a large volume of data from different data sets and process it over time for each component. In our study we will focus on analyzing each component's history to model the exact state of the component after all the events that occurred in the past. We provide continious observation of each component over time, and use its history to predict its performance in the future.
- Third, there is a lack of dynamic component-specific asset management capabilities that could improve the overall performance of the system. The optimization algorithms used in earlier works did not have access to a precise data-driven prediction of asset

states. We provide an optimal maintenance scheduler that uses the outputs of a predictive risk analysis to determine the best set of actions for each individual component in the network.

- One of the most important aspects that is missing in the literature is the capability to use spatial and temporal interdependancies between components and variables to provide more accurate prediction. The spatiotemporal correlation of such data sets is a complex task. In addition, after the correlation, the ability to draw knowledge from spatiotemporal interdependencies of the components and events poses an additional challenge when developing a prediction algorithm. The inclusion of spatiotemporal observation of variable interdependancies can significantly improve the prediction accuracy. In addition, such a model is inherently robust to bad and missing data.

CHAPTER III

PROPOSED RESEARCH APPROACH [*]

3.1 Introduction

The proposed research approach is built to target the three main limitations of existing solutions that were described in Chapter II: 1) lack of extensive data sets and methods for their integration and spatiotemporal correlation, 2) limited predictive capabilities that are not targeting every component individually, and 3) lack of prediction-based optimal maintenance strategies that could reduce the costs of asset management dynamically over different time horizons. To solve these problems and develop a unified prediction framework that can assess many different applications of weather impacts to power system transmission and distribution, we propose a three-level approach that will be described in section 3.3.

3.2 Hypothesis and Objectives

The ***hypothesis*** of the research is that more accurate predictions are possible by structured learning from merged heterogeneous Big Data. The hypothesis will be validated by developing ***Optimized Risk-Based Asset Management*** capable of assessing equipment deterioration continuously across space and time, leading to an improved on-demand maintenance strategy. The main objectives of this study are:

1) Implement short- and long-term prediction of outage occurrence probabilities.
2) Develop dynamic optimal asset management strategies based on outage prediction.
3) Demonstrate the capabilities of the predictive framework on multiple applications.

---

[*] This section is in part a reprint with permission of the material in the following paper: T. Dokic, M. Kezunovic, "Optimized Asset Management in Distribution Systems Based on Predictive Risk Analysis," Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion - MEDPOWER, Dubrovnik, Croatia, November 2018.

3.3 The Three-Level Approach

The goal is to improve the current asset management practices on three levels:

1 *Data Management*: The study includes a variety of data from multiple sources. Some of the challenges of collecting the diverse data sets are: 1) data storage: large volumes of data need to be collected and preprocessed in a way that increases the overall value of data for the utility; 2) data integration: the data is collected at multiple temporal and spatial scales; 3) data quality: the data sets may contain bad and missing data, and the uncertainty levels of data varies from one set to another. In this dissertation we show how such cases may be handled. The expected benefits include the generation of a unified spatiotemporally-referenced database capable of serving multiple applications, and the capability to provide accurate prediction even with bad and missing data, or data with high uncertainty.

2 *Data Analytics:* The study uses prediction algorithms [93-96] capable of leveraging the spatial and temporal aspects of heterogeneous data as a knowledge source. Graph-based machine learning methods are used for prediction. We demonstrate the following features of the proposed data analytics: 1) high accuracy of prediction based on modeling of spatial and temporal interdependencies between variables, 2) capability to provide prediction on multiple temporal and spatial scales, 3) robustness to missing and bad data thanks to the use of variables from the nearby nodes. This kind of system is capable of serving multiple departments responsible for power distribution applications: 1) the prediction of expected locations of outages can be used by operators to make better decisions; 2) asset management can benefit from more precise component-oriented prediction of asset failures that relies on spatial interdependencies

of components; 3) outage management can use the results to provide better outage location identification and faster restoration practices.

3 *Economic Impact:* Maintenance decision-making is focused on minimizing the risk level while maintaining economic investment limits. We assume that while the cost of periodic maintenance stays the same, the reactive maintenance cost can be optimized and reduced. The main benefits of the economic impact analysis are: 1) the capability to define user-specific optimization problems for any weather-related application in the network; 2) risk-driven optimization based on prediction of future events provides a new framework for dynamic allocation of funds and resources; 3) the value of the Big Data is increased by careful balancing of the costs of method implementation and asset management expenses on one side, and reduction of reactive maintenance costs and asset failures on another side.

The above approaches result in a novel asset management that enables the following capabilities:

- Assessing equipment deterioration continuously across space and time by learning from heterogeneous data
- Performing real-time risk assessment on multiple temporal and spatial scales by assessing the hazards and vulnerabilities
- Developing an optimal asset management strategy that reduces the outage risk

The goal is to integrate the environmental data into the power system models and studies, build a model that integrates and exploits all types of data, implement data analytics that evaluate system and component risk in real time, and contrast the existing static asset management practices with the new dynamic approach. The three-level approach targets each one of these goals with an appropriate action.

3.4 Conclusion

We have postulated the hypothesis and provided the approach for assessing its validity. Our main premise is that the high accuracy of prediction will be achieved for all cases even with a large high percentage of bad and missing data, and that such results can be used for development of optimal asset maintenance strategies that will significantly improve the reliability of the system.

How the three-level approach is implemented will be described in Chapters V, VI, and VII. The details of implementations for two different applications will be described in Chapters VIII and IX. We will demonstrate how this kind of approach can lead to the improvement of system performances under severe weather conditions for multiple instances of weather impacts. We will illustrate how the proposed data analytics enable one to optimize the use of different assets in the network, on both the transmission and distribution sides, while reducing the cost.

CHAPTER IV

ASSET MANAGEMENT*

4.1 Introduction

Current practices use several approaches to asset maintenance scheduling [19,97]:

- Run-to-Failure (RTF): in this approach maintenance is only performed after the component fails. The advantage is that there are no expenses associated with equipment monitoring and analysis, but the disadvantage is that there is no mechanism to predict outages due to the equipment failure resulting in higher cost to reinstate service.
- Condition-Based Monitoring (CBM): this type of maintenance is initiated by the monitoring equipment indicating that certain performance degradation thresholds are exceeded, requiring maintenance action. While this method allows prevention of an outage by "just in time" action, it is typically costly due to the requirement that each individual component is closely monitored.
- Reliability-Centered Maintenance (RCM): the maintenance schedule is prioritized based on the likelihood of equipment failure. This kind of maintenance scheduling does observe the whole system and prioritizes the maintenance area. However, in the existing RCM studies, the likelihood is determined statistically and it is equal for all components, neglecting the variety of factors affecting the individual components over

---

* This section is in part a reprint with permission of the material in the following papers: (1) M. Kezunovic, T. Dokic, "Predictive Asset Management Under Weather Impacts Using Big Data, Spatiotemporal Data Analytics and Risk Based Decision-Making," 10th Bulk Power Systems Dynamics and Control Symposium – IREP'2017, Espinho, Portugal, August 2017. (2) T. Dokic, M. Kezunovic, "Predictive Risk Management for Dynamic Tree Trimming Scheduling for Distribution Networks," IEEE Transactions on Smart Grid, September 2018. Copyright 2018, IEEE. (3) T. Dokic, M. Kezunovic, "Optimized Asset Management in Distribution Systems Based on Predictive Risk Analysis," Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion - MEDPOWER, Dubrovnik, Croatia, November 2018.

time, which prevents the development of optimal spatiotemporal maintenance strategies.

- Optimization Techniques (OT): the maintenance schedule is optimized based on the economic impacts. This kind of maintenance considers restrictions such as availability of maintenance crews, travel expenses, and restricted time intervals, but still does not get the benefit of the predictive risk assessment based on unfolding threats and vulnerabilities.

The approach proposed in this dissertation assures that several shortcomings of the previous methods listed in Section 2.2 are addressed. The overview of characteristics of conventional asset management approaches and our proposed method is presented in Table 2.

4.2 Asset Management for Vegetation

This section describes the mechanisms of weather and vegetation impacts on vegetation-caused outages, and current vegetation management practices implemented by the utilities [53]. As presented in Fig. 4, there are two major classes of vegetation-related feeder outages in distribution systems. They are differentiated by the tree coming in contact with feeders due to 1) overgrowing the feeder height, and 2) being forced into a contact with the feeder due to wind or some other similar weather impacts.

Starting from the most recent tree trimming performed, the vegetation-caused failure probability is constantly increasing [98]. For predicting the potential of vegetation to cause faults subsequent to the last tree trimming, the most important factor is the vegetation canopy growth

Table 2 Comparison of asset management approaches, reprinted from [99]

Approach/Feature	Run-to-failure/Periodic	Condition-based	Reliability-centered	Optimization techniques	Our Approach
Monitoring cost	No expenses	High	High	High	High
Cost of reinstating services	High	Low	Low	Low	Low
Preventive capability	No	Yes	Yes	Yes	Yes
System or component level	Component level	Component level	System level	Both	Both
Data	No	One or several different measurements	One or several parameters observed	One or several parameters observed	Big Data – wide variety of parameters
Predictive	No	No	Yes – statistical	No	Yes – better accuracy with machine learning
Spatiotemporal analysis	No	No	Limited	No	All data spatiotemporally referenced
Dynamic real-time assessment	No	Yes	Limited	Limited	Yes
Interdependencies between components	No	No	No	No	Geographical and electrical

rate. There are two types of models for estimating the canopy growth dynamics [100]: 1) process-based models that aim at defining the processes that cause tree growth [101], and 2) empirical data-based models [102]. The maximum tree crown spread represents the maximum width of the tree crown (branches, leaves) along any axis. It is affected by the tree's age, last tree trimming date, application of herbicides or growth regulators, and weather impacts (primarily temperature and precipitation) [100].

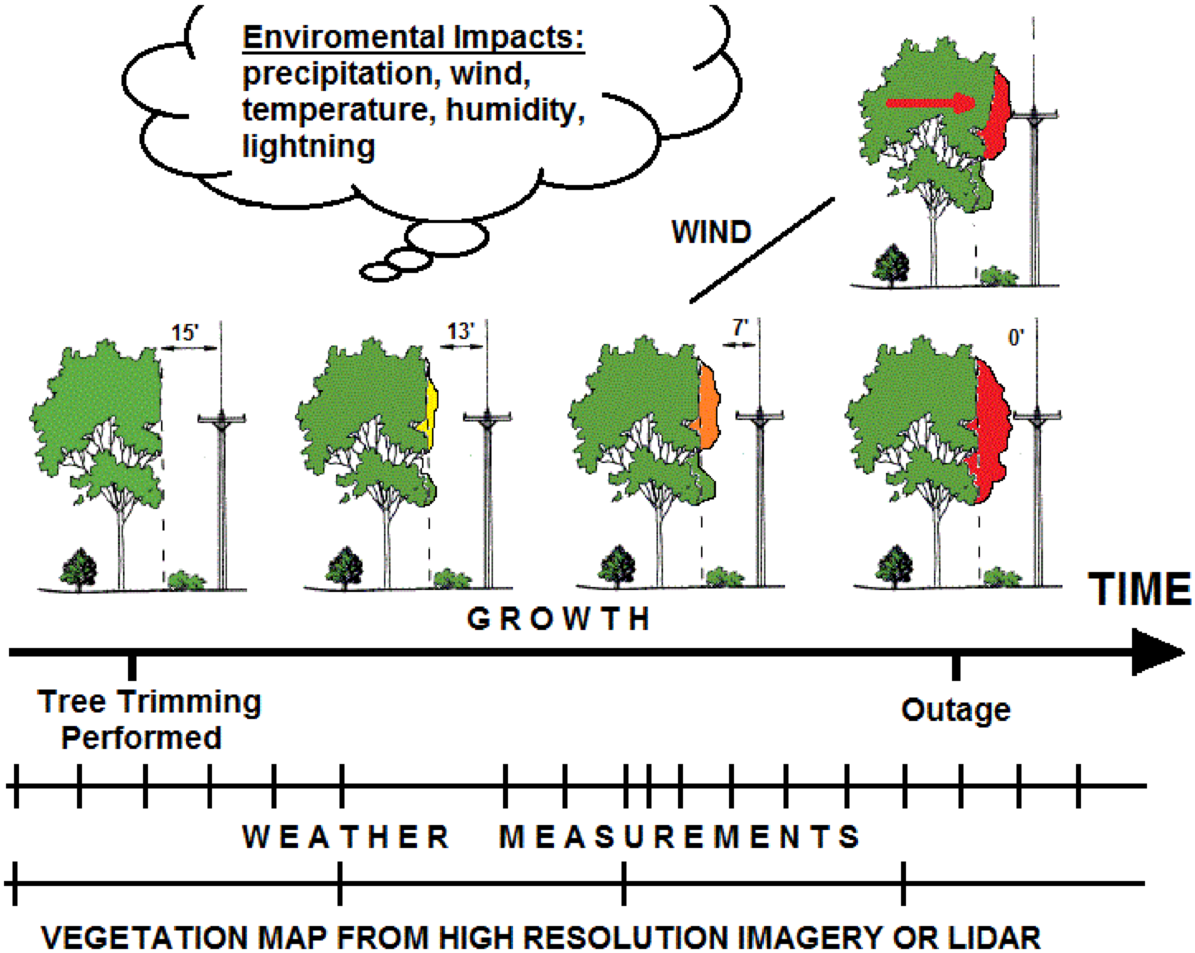

Figure 4 Environmental impact on vegetation management, reprinted from [5]

The measured electrical behavior and physical processes and effects surrounding the vegetation-related faults were described in detail in [103, 104]. It was concluded from the experimental results that while the initial current during the tree contact can be quite low (~1A), after a complete carbonization path in the tree branch is formed, the current magnitude quickly increases to a much higher level.

The weather parameters that can affect vegetation-related outages are wind speed, direction, and gusts, precipitation, temperature, humidity, pressure, and lightning, as listed in Fig.

4. The impact of high-speed wind and heavy precipitation may cause trees to come into contact with distribution feeders due to the following reasons: a) branches break off and fly into lines, and b) complete trees topple when moved by wind [105]. The temperature, precipitation level, and humidity have impacts on the tree growth rate. In combination with the type of soil, they are the main factors dictating a tree's growth rate.

Vegetation maintenance staff are in charge of maintaining the feeder clearance to the surrounding vegetation. This includes trimming and removal of trees around the distribution poles and lines. Distribution lines are often placed near the surrounding vegetation due to relaxed right-of-way requirements. Due to the high expenses of trimming large areas populated by many distribution feeders, it is not economical to have all trees securely trimmed at all times, so a more economical trimming schedule is needed.

In most cases, the process of tree trimming is applied by utilities based on a predetermined periodic schedule. Each feeder section is given a tree trimming frequency, e.g. three or five years, based on the operating voltage and required clearance, leading to the standard fixed interval schedule [106]. The only other occasion when the schedule would be changed is as a reactive measure to a vegetation-caused outage. There are two types of reactive measures that can be distinguished: 1) only the faulted area is maintained, and 2) the entire tree trimming zone is trimmed. In addition to tree trimming, some utilities inject growth-retarding chemicals into trees (tree-growth regulators) or apply herbicides [106].

The current maintenance practice relies on a visual inspection by staff using helicopters, airplanes, ground vehicles, or people walking up to the lines [107]. Because of the high cost of this practice, it is of economic benefit to develop visual inspection methods that can provide automatic identification of dangerous zones. If a predictive method as proposed in this dissertation is used, several benefits are expected:

- In contrast to a limited amount and variety of data used in literature [54, 57, 59] in our study all of the aspects of vegetation related outage events are observed over time by collecting a variety of measurements surrounding every outage event in the network.
- In contrast to studies [54,57], spatiotemporal predictive methods will enable accurate mapping of vegetation outages in the distribution network.
- The processes of data collection, knowledge extraction, prediction, and optimization are all automatic, without any need for manual inspections in distribution, which was only possible in transmission in the past [61, 62].
- For the first time an optimal tree trimming schedule will be implemented based on the predictive risk analysis.

4.3 Asset Management for Insulators

The insulators play an important role in power transmission system, as the integrity of an overhead transmission line is directly governed by their electrical and mechanical performance [19]. Statistically, while insulators account for only 5% to 8% of the direct capital cost of the transmission line, more than 70% of the line outages and up to 50% of line maintenance costs are caused by the insulator-induced outages [21].

Insulation coordination is the study used to select insulation strength to withstand the expected stress caused by lightning and switching overvoltages. There are two approaches to tackle this problem: deterministic and probabilistic [97]. In the former, the minimum strength is set to be equal to maximum stress. In the latter, the lightning flashover rate and lightning-related failures are calculated statistically, and insulation strength is selected accordingly.

Insulation strength can be described using the concept of BIL [97]. Statistical BIL represents a voltage level for which the insulation has a 90% probability of withstanding and 10% probability of failing. Standard BIL is expressed for a specific wave shape of lightning impulse and standard atmospheric conditions. A typical lightning impulse wave shape is presented in Fig. 5 [108]. Conventionally, BIL is determined by the manufacturer by performing a set of tests at standard atmospheric conditions [97]. It should be noted that these tests are performed before any kind of environmental exposure of the insulator, so they may not reflect the actual strength of the insulator after the prolonged exposure.

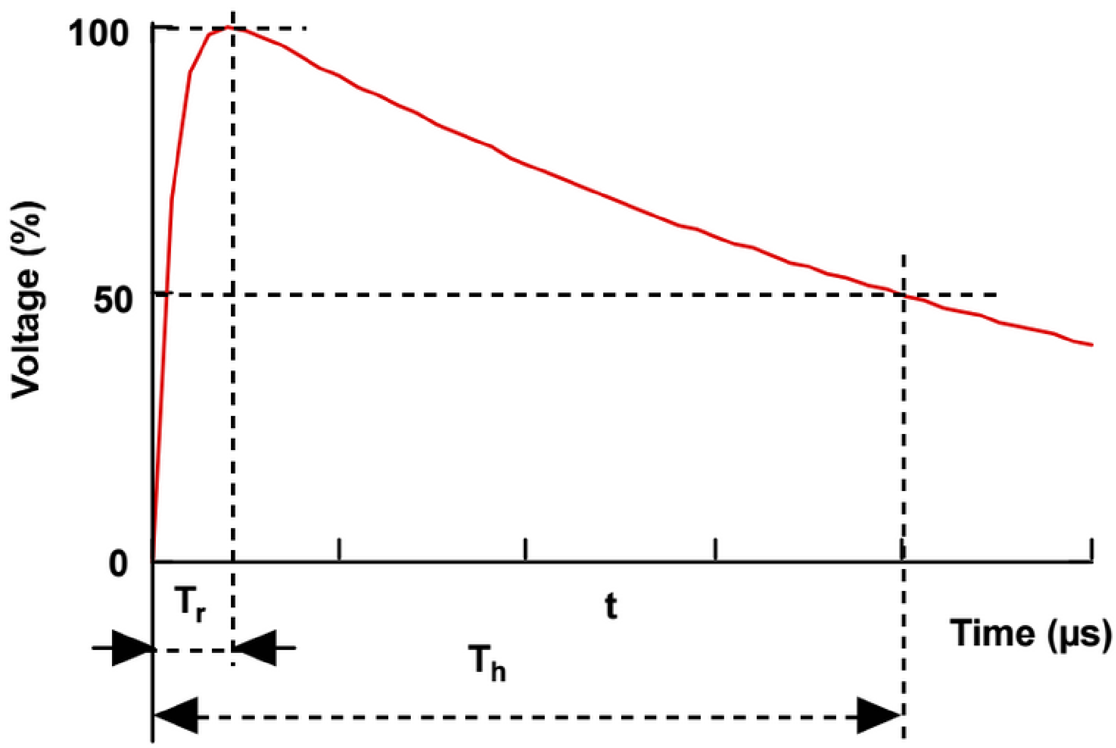

Figure 5 Standard lightning impulse (T_r = 0.1–20 μs, T_h < 300 μs, where T_r is the time-to-crest value, T_h is the time-to-half value), reprinted from [108]

Insulators exhibit two types of failures, [109]: 1) mechanical failures caused by physical deformities due to manufacturing defects or severe material erosion; and 2) electrical stress failures caused by increased leakage current, mostly due to a high number of experienced flashovers. Due to exposure to different environmental impacts, the mechanical and electrical performances of insulators deteriorate over time [109]. These changes in insulator performances are not always easily observable.

Insulator deterioration can be classified into two stages, [110]: 1) the deterioration of hydrophobic properties where an insulator may age chemically, but it still retains its electrical properties; 2) hydrophobic properties of an insulator start to deteriorate causing the degradation in insulator electrical performance. Based on the study presented in [109], the second stage can be further separated into three groups: i) *weathered*, with a small or moderate loss of hydrophobic properties; ii) *mature*, with a very low hydrophobicity; iii) *at risk*, with a fully hydrophilic surface, or total loss of insulation properties. An overview of the deterioration rates is presented in Fig 6 [111].

There are multiple measurements that can be performed to estimate the conditions of network insulators [109]. At the network level, the history of outages and disturbances can be used to quantify the insulator failure rates. At the component level, the individual insulator can be tested for its electrical and mechanical properties. The tests can be destructive (only performed in laboratory) or non-destructive (performed in field with system energized or not energized depending on the test type) [111]. The following parameters can be measured in a nondestructive way [109]: i) leakage current magnitude, ii) flashover voltage, iii) electric field distribution, iv) corona discharge, v) radio interference voltage. In addition, it is possible to characterize the

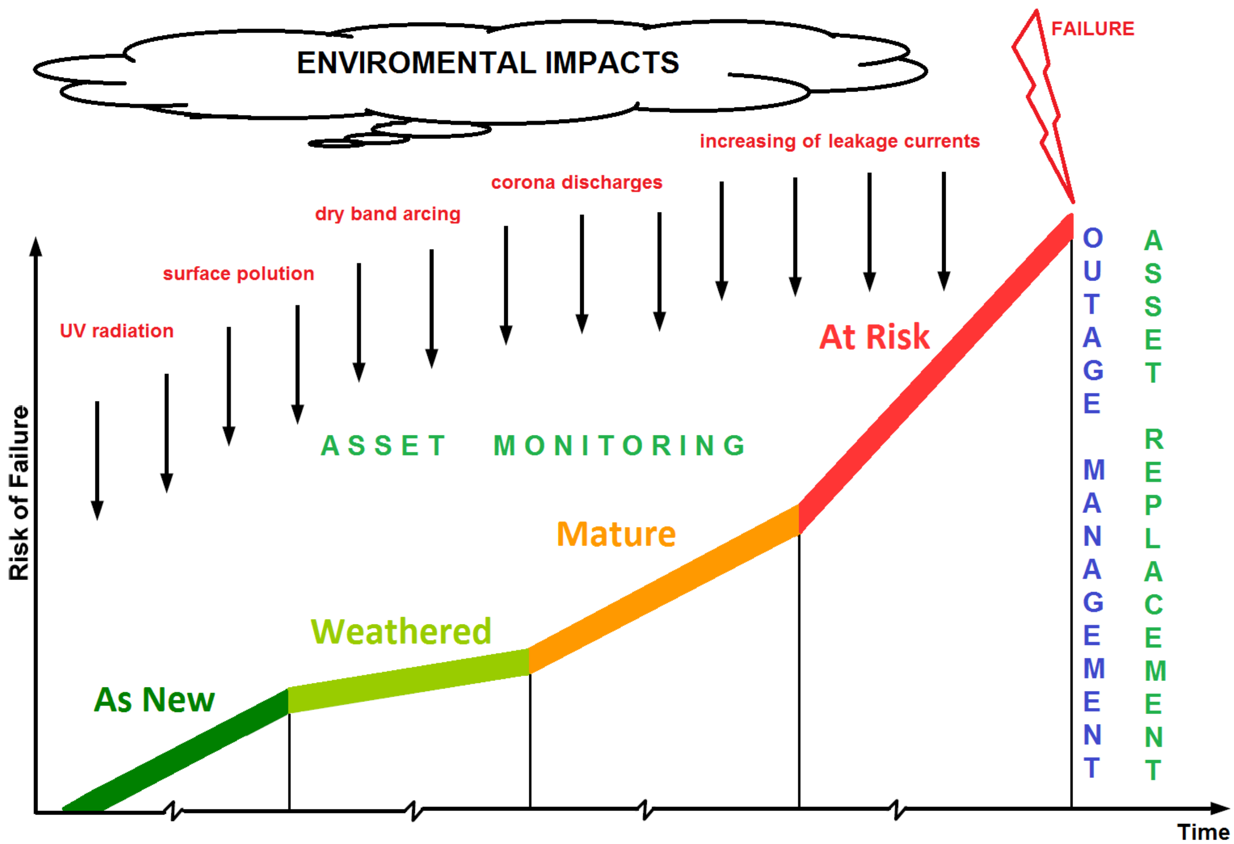

Figure 6 Insulator deterioration process due to environmental impacts, adapted from [111]

insulator material by performing one of the following in-field tests [109]: i) visual inspection, ii) infrared reflection thermography, iii) hydrophobicity, and iv) remote chemical analysis.

This research focuses on the deterioration of the electrical performance of insulators during the second stage of deterioration when the insulator is experiencing a loss of electrical strength. During this period, the manufacturer's BIL no longer can be used as the measure of insulator electrical strength. Instead, the BIL curve can be updated after every event in the network to reflect the changes in insulator performance caused by deterioration. This provides a more realistic approach to insulation coordination instead of simply using the initial manufacturer's BIL, and enables dynamic component maintenance scheduling. With the capability to precisely estimate the current state of insulator at any moment in time, the prediction algorithm can rely on a more precise spatiotemporal set of input data describing events of interest.

Overhead line insulators are exposed to variety of environmental impacts, [19]: i) lightning strikes, ii) temperature and pressure variations, iii) ultraviolet radiation and ozone, iv) wind impact, v) rain, humidity, hail, snow, fog, and vi) pollution. In addition, a variety of environmental factors affects the probability and characteristics of flashover. Vegetation coverage around the line will lower the probability of a lightning strike affecting the network, the phenomenon called "shielding by trees" [112]. Elevation data is of importance also, since lightning strikes are more likely to affect locations with higher altitude [113]. The type of soil at the tower location determines the tower grounding resistance, which has a big impact on overvoltage propagation on the line [114].

To estimate the expected stress on insulation, the failure risk is calculated statistically as (Fig. 7):

$$R = \int_0^\infty f(V) \cdot D(V) dV \tag{1}$$

where $f(V)$ is the probability of overvoltage occurrence and $D(V)$ is the probability of a disruptive discharge. The probability of an overvoltage occurrence can be described with a density function as follows:

$$f(V) = \frac{1}{\sigma_O \sqrt{2\pi}} e^{-\frac{(V-V_O)^2}{2\sigma^2}} \tag{2}$$

where V_O is the voltage for which probability density of overvoltage occurrences has a maximum, and σ_O is the standard deviation. The probability of a disruptive discharge can be expressed with a cumulative function:

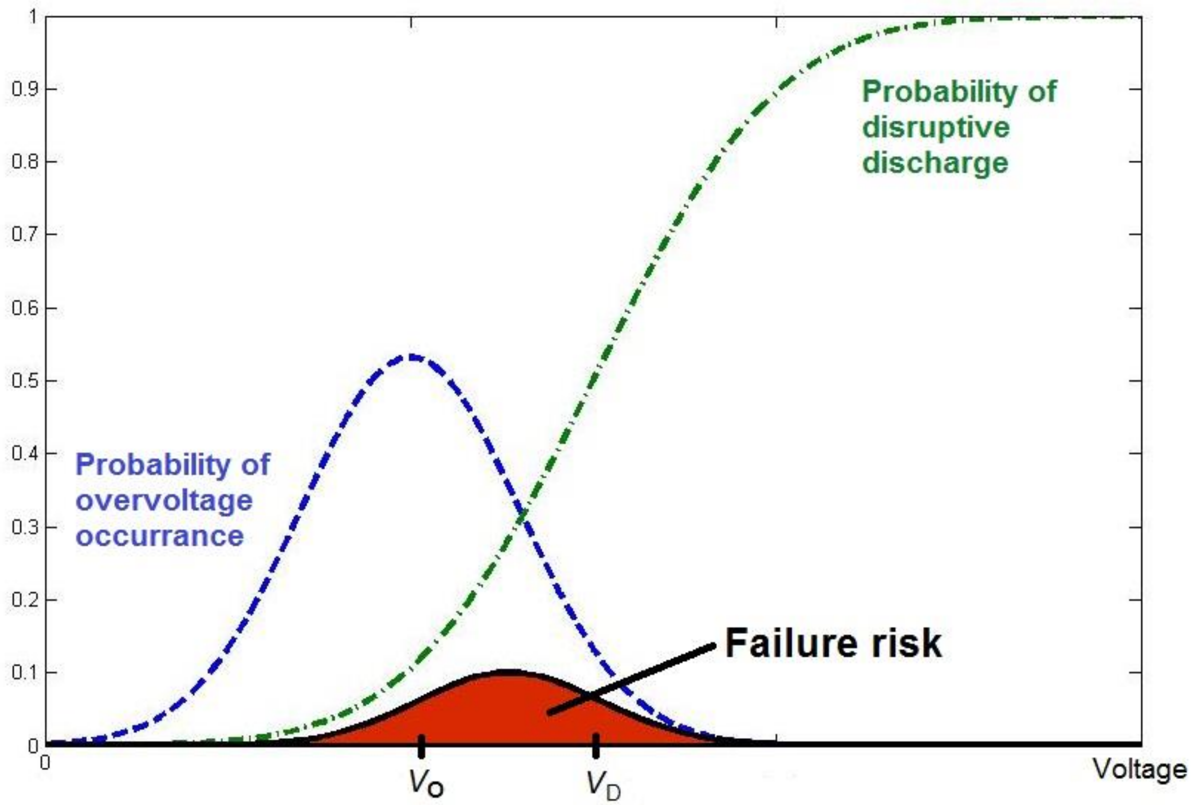

Figure 7 Risk of component failure, adapted from [115]

$$D(V) = \frac{1}{\sigma_D\sqrt{2\pi}} \int_{-\infty}^{V} e^{-\frac{(V-V_D)^2}{2\sigma^2}}\, dV \tag{3}$$

where V_D is voltage for which the insulation has a 50% probability of a flashover, and σ_D is a standard deviation. More details about probabilistic models of insulation flashover can be found in [115-118].

Weather data is used to calculate *BIL* under nonstandard atmospheric conditions [97], BIL_A as:

$$BIL_A = \delta\ H_C BIL_S \tag{4}$$

where BIL_A is the *BIL* under nonstandard conditions, BIL_S is the standard *BIL*, δ is the relative air density, and H_C is the humidity correction factor. Relative air density can be calculated using:

$$\delta = \frac{PT_S}{P_S T} \tag{5}$$

where T_S and P_S are standard temperature and pressure respectively; T and P are measured temperature and pressure respectively. The humidity correction factor is equal to 1 for rainy conditions and for dry conditions can be calculated using:

$$H_C = 1 + 0.0096 \cdot \left[\frac{H}{\delta} - 11 \right] \tag{6}$$

where H is humidity.

If the predictive method proposed in this dissertation is used, several benefits are expected:

- The study will enable prediction of lightning performance for each individual component based on all the historical data collected for outage events, which was not enabled in any of the existing studies [66-74].
- A system for real-time tracking of weather impacts and atmospheric conditions will be provided. This will extend the inclusion of weather parameters from [73] with the capability to use this data for prediction of future states of insulators.
- Unlike in conventional studies that use constant insulator strength over time [66-74], the insulator strength will be updated over time based on the experienced lightning events and atmospheric conditions around the network.

- For the first time, optimal maintenance, replacement, and placement strategies will be enabled based on predictive risk mapping.

4.4 Conclusion

The review of the literature identifies several common shortcomings of the existing methods:

- Limited data is used, including only utility data and limited sources of weather data
- No precise tracking of component-related weather impacts; instead, the weather impacts are averaged over time or over larger area
- Very limited predictive capabilities, where spatiotemporal interdependencies between component and events are not explored as a source of knowledge
- No optimal asset management strategies based on predictive methods

Compared to the mentioned methods, we propose a three-level approach that has several advantages: 1) it introduces the capability to process, utilize, and visualize larger amounts of diverse data; 2) it enables implementation of predictive analytics using spatiotemporal data where spatial interdependencies between components are considered; 3) it enables dynamic maintenance scheduling based on real-time observation of network components' states and surrounding conditions. All of this can lead to improvement in practices for vegetation-related asset management and lightning equipment asset management because it enables real-time tracking of weather impacts, the prediction of expected vulnerabilities in the future, and development of optimal decision-making strategy for mitigation of such severe weather impacts. Further details are described in Sections 4.2 and 4.3. In Chapters V, VI, and VII we will describe how this can be achieved following the three-level approach.

CHAPTER V

DATA MANAGEMENT*

5.1 Introduction

This chapter describes the first level of the framework, namely Data Management. Here we will list different datasets that were used in the study and discuss the kind of processing that was performed to prepare the data for the second level of the framework, namely data analytics. The correlation between the choices of data sources and approaches to data correlation are tied to the proposed data analytics and the reasoning is explained.

We are considering many different data sources, some coming from electric utilities, such as historical outage data, asset data, network GIS data, etc. Other data come from other sources, such us weather data, vegetation data, lightning detection network data, etc. Correlating such data offers many benefits:

- Building a unified spatially and temporally referenced database that can serve a variety of weather impact applications in power systems. For example, various applications need weather forecast data, and may use different parameters collected by weather forecast models. Vegetation indices can be used for prediction of vegetation-related outages, but they can also be used to identify areas shielded by trees for lightning protection studies. A unified framework takes care of all the preprocessing needs to extract the parameters needed for every application. However, the concentrated storage

---

* This section is in part a reprint with permission of the material in the following papers: (1) M. Kezunovic, T. Dokic, "Predictive Asset Management Under Weather Impacts Using Big Data, Spatiotemporal Data Analytics and Risk Based Decision-Making," 10th Bulk Power Systems Dynamics and Control Symposium – IREP'2017, Espinho, Portugal, August 2017. (2) T. Dokic, M. Kezunovic, "Predictive Risk Management for Dynamic Tree Trimming Scheduling for Distribution Networks," IEEE Transactions on Smart Grid, September 2018. Copyright 2018, IEEE.

of data and processing routines increases the value of this data, since it is not replicated for every application separately.

- Identification of event characteristics by spatiotemporal correlation of all the measurements that surround it. For example, for one outage on a specific location, we want to be able to quickly extract various parameters that are not necessarily measured at that location (but in the close vicinity), or at that moment in time (but soon after or soon before). Weather measurements for an event may come from a set of stations that are closest to the outage location, and measured in time that is within one minute of an outage.
- Real-time event tracking and event prediction. This aspect describes the capability of the framework to efficiently process all the data and provide answers within a specified time limit.

The initial step is to collect all of the datasets and prepare them for the use in the proposed data analytics, which includes removing all the unnecessary data, as well as imposing quality control by removing data that may be bad or corrupted. After that, the specific set of parameters of interest for the approach we are using can be extracted from the datasets. The most important step is to implement spatiotemporal correlation between all datasets to make sure all the parameters are associated with the specific time of the event at the specific location. This is a necessary step for the development of prediction algorithm inputs and outputs required for training. For each historical event we want to analyze, the prediction algorithm needs to have a single time step and location for each component and its associated measurements. For example, if there was a lightning caused outage on a distribution line, we want to collect all the weather measurements and vegetation indices at the specific location of this outage at the time the event occurred.

5.2 Data Sources

Tables 3 and 4 contain the list of data sources and their characteristics. The first dataset, ASOS, contains historical weather measurements. The measurement stations are located sparsely over the area of interest. Fig. 8 shows locations of weather stations in the analyzed area, and Table 5 presents an example of one weather data point. The NDFD contains weather forecasts (current and archived). Fig. 9 shows an example of the weather forecast map for wind speed. The specific selection of data assures that the following aspects of the weather impacts are captured:

- Historical weather measurements are collected with high temporal resolution and spatially interpolated to the component locations. As presented in Table 3, historical weather measurements coming from ASOS land-based weather stations come with a maximum time resolution of 1 min. These weather stations are sparsely located (there are about 900 stations across the US), and need to be spatially interpolated to the exact location of an event.
- Real-time weather forecast is collected at each location and used as the temporal reference for the prediction algorithm output time referencing. As presented in Table 3, the temporal resolution of weather forecast data coming from NDFD is 3 hours (except for precipitation, which is collected with a 12-hour resolution). Because the new weather forecast becomes available every 3 hours, this is chosen as the time horizon for the prediction algorithm. Every three hours we generate the new set of risk maps based on the most recent weather forecast.

A number of high-resolution imagery files are available on TNRIS as part of NAIP, Fig 10. The TPWD EMST dataset contains the classification of vegetation types by area. An example of a simplified classification map is presented in Fig. 11. The NASA 3D Global vegetation map is presented in Fig 12. This map is created by combining radar and LIDAR remote sensing.

Table 3 Non-utility Data Sources and Characteristics, reprinted from [10]

	Source	**Data Type**	**Temporal Coverage**	**Spatial Coverage**	**Temporal Resolution**	**Spatial Resolution**	**Measurements**
WEATHER	ASOS [30]	Land-Based	2000-Present Used: 2011-2015	USA	1 min	900 stations	Air Temperature, Dew Point, Relative Humidity, Wind Direction, Speed and Gust, Sea Level Pressure, Sky, Precipitation
	Vaisala NLDN [31]	Lightning Data	1989-Present Used: 2011-2016	USA	Instantaneous	Median Location Accuracy <200m	Date and Time, Latitude and Longitude, Peak amplitude, Polarity, Type of event: Cloud or Cloud to Ground
	NDFD [32]	Weather Forecast Data	Present – 7 days into future Used: 2016	USA	3 hours	5 km	Wind Speed, Direction, and Gust, Temperature, Relative Humidity, Tornado Probability, Probability of Severe Thunderstorms, etc.
VEGETATION	TPWD [33]	EMST	2015	Texas	static	10 m	Distribution of different tree spices
	TNRIS [34]	NAIP	2010, 2012, 2014, 2015, 2016, 2017	Texas	year	50 cm – 1 m	High Resolution Imagery
	NASA [119]	3D Global Vegetation Map	2011	World	static	1 km	Canopy height data

Table 4 Utility Data Sources and Characteristics

Data Type	Temporal Coverage	Spatial Coverage	Temporal Resolution	Spatial Resolution	Measurements
Historical Outage Data	2011-2016	Utility area	instantaneous	Feeder section	Location, start and end time and date, number of customers affected, cause code
Tree Trimming Data	2011-2016	Utility area	day	Feeder	Feeder location, date, trimming period, number of customers affected, cost of trimming
Network GIS data	2016	Utility area	static	Infinity (shapefile)	Poles: location, material/class, height Feeders: location; conductor size, count, and material; nominal voltage
Historical Maintenance Data	2011-2016	Utility area	day	Tower location	Start and end date and time, location, type (maintenance, replacement), cost, number of customers affected
Insulator asset data	2016	Utility area	static	Infinity (shapefile)	Surge Impedances of Towers and Ground Wires, Footing Resistance, Component BIL
In-field measurements	2011-2016	Utility area	instantaneous	Tower location	Leakage Current Magnitude, Flashover Voltage, Electric Field Distribution, Corona Discharge Detection, Infrared Reflection Thermography, Visual Inspection

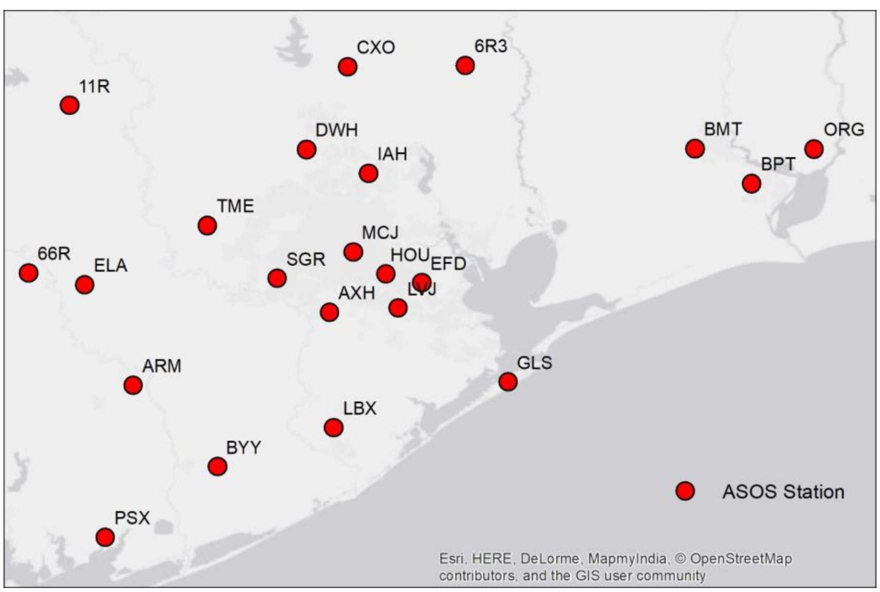

Figure 8 ASOS weather stations locations [30]

Table 5 ASOS weather station measurement example

date/time	tmpf	dwpf	relh	drct	sped	alti	mslp	p01i
9/9/2018 0:53	80.1	78.1	93.64	350	11.5	29.84	1010.3	0.04
vsby	**gust_mph**	**skyc1**	**skyc2**	**skyc3**	**skyl1**	**skyl2**	**skyl3**	**wxcodes**
7	M	FEW	SCT	BKN	1000	3500	5000	-TSRA

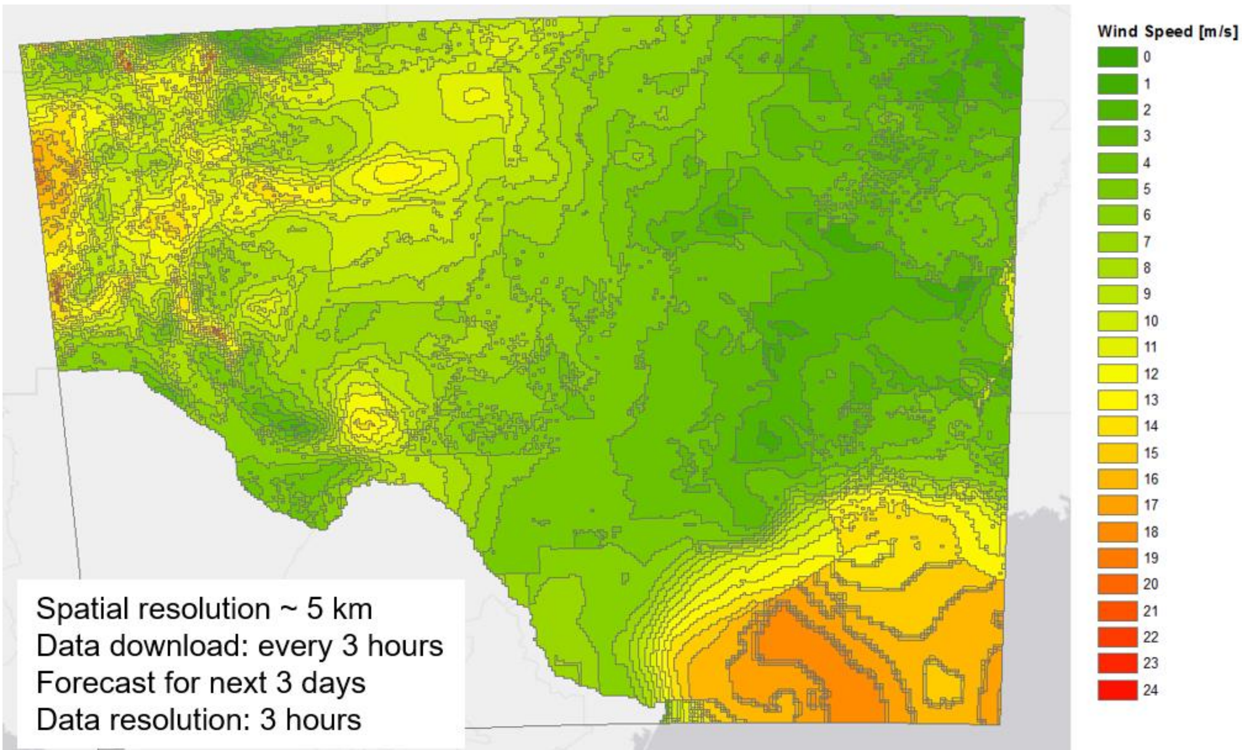

Figure 9 NDFD map example for wind speed [32]

5.3 Spatiotemporal Correlation of Data

Geographic Information System (GIS) and Global Positioning System (GPS) together provide a framework for conducting spatiotemporal correlation between the weather threats and their corresponding impacts. GIS provides a platform for spatial integration of various datasets.

Figure 10 NAIP Imagery example [34]

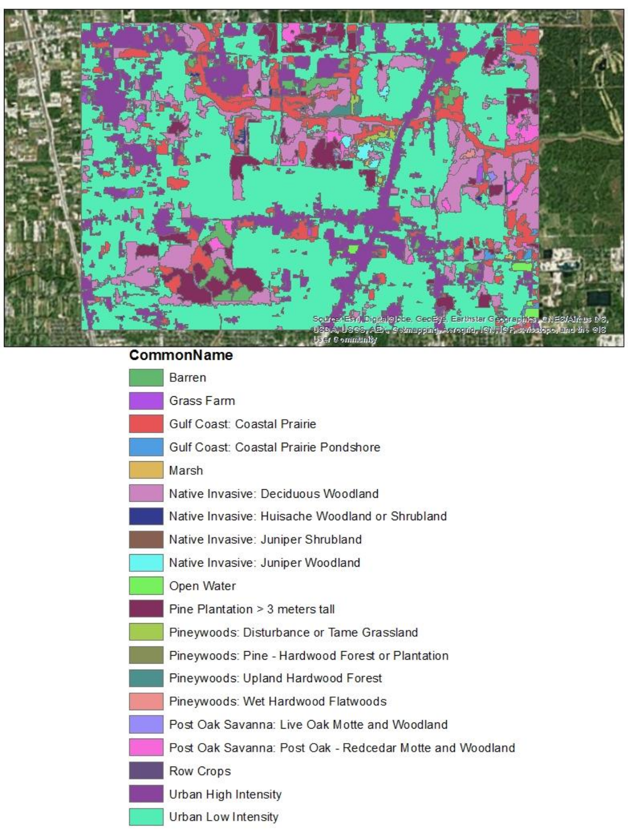

Figure 11 EMST data example [32]

The GPS clock is used to temporally reference events and measurement points. As stated in [120], the spatial and temporal correlation of data plays an essential role in the process of integrating Big Data analytics into the electric power industry applications, since it enables the identification of measured values of various parameters at the time of an event at the location of interest. Spatial correlation of data is done by integrating different data sets as layers of GIS, while GPS is used for time synchronization between events, and for synchronizing the sampling. All of the spatial processing of the data in this dissertation is done using ESRI ArcGIS [121]. Temporal data processing is done using the Python standard library *datetime* [122].

The spatiotemporal correlation of diverse data makes sure that every dataset has associated unique spatial and temporal references, and determines spatial and temporal relationships between different datasets. For example, for every historical vegetation-caused outage, we want to know the measured and forecasted weather conditions at that specific location, the distance between the feeder and the tree, etc. All these parameters are used as inputs to the prediction algorithm described in Chapter VI.

5.3.1 Spatial data analytics

As reported in [19], two distinct categories of GIS data, spatial and attribute data, can be identified [123]. Data which describes the absolute and relative context of geographic features is spatial data. For transmission towers, as an example, the exact spatial coordinates are usually accessible by the operator. To provide additional characteristics of spatial features, attribute data is included. Attribute data includes all other characteristics that can be either quantitative or

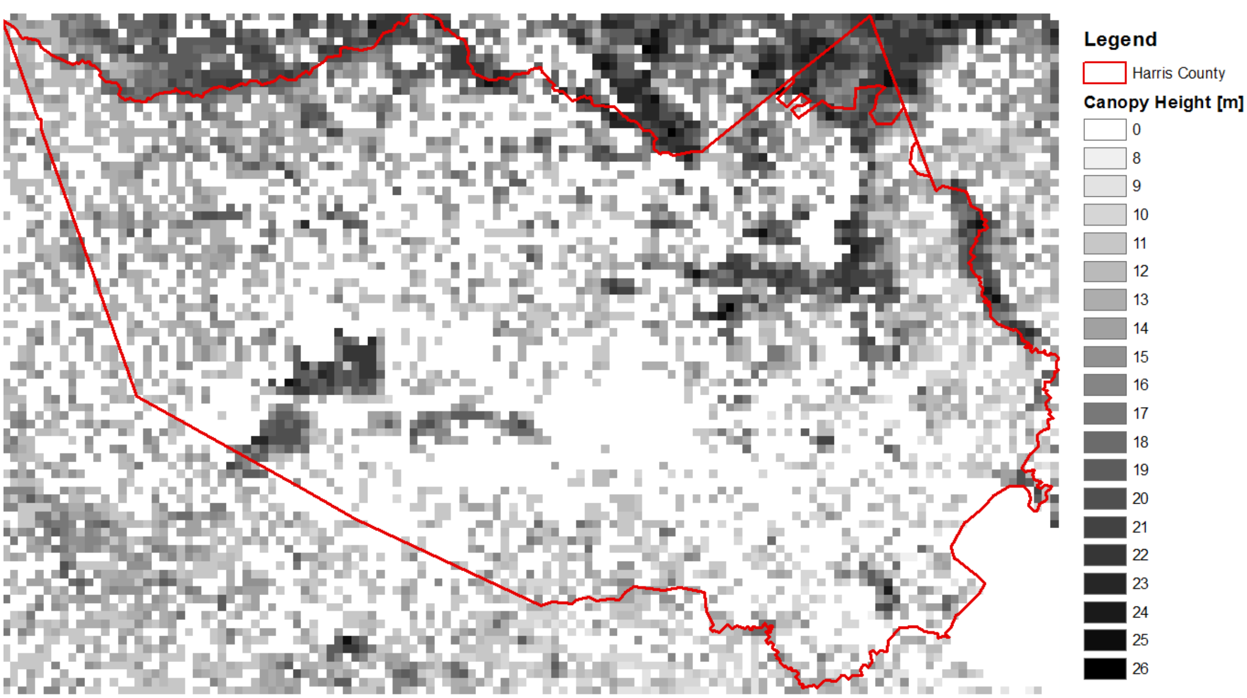

Figure 12 3D Global Vegetation Map [119]

qualitative. For example, a table with the physical characteristics of a transmission tower can be described with the attribute data. Any kind of data with a spatial component can be integrated into GIS as another layer of information, [121]. As new information is gathered, these layers can be automatically updated.

In terms of spatial data representation, raster and vector data can be used. In the case of vector data, polygons, lines and points are used to form shapes on the map. Raster data presents information as a grid, where every cell is associated with one data class. Typically, different data sources will provide different data formats and types. Some of the data received in tables with geographical coordinates is converted into GIS shapefiles that contain both geographic references and attribute tables for every dataset. Some of the data, such as satellite imagery, comes into the geodatabase as raster files. Different tools need to be used within GIS to extract needed parameters from raster files and shapefiles for the prediction algorithm. Also, the type of a file dictates the spatial resolution; while shapefiles have a precise location of the point/line/polygon, raster files average the measured value within a grid, where the size of the grid's cell dictates the spatial resolution of data.

To deal with different spatial resolutions of data, we use multiple approaches. We use spatial interpolation to increase the spatial resolution of datasets when such action is required. For example, ASOS weather station measurements are only available in certain sparse locations. To improve the spatial resolution of ASOS data, we will use interpolation to estimate weather parameters for the rest of the network area. We will also use spatial joins to project one dataset to a different dataset location. For example, we will need to extract all the NDFD data associated with a specific network zone. It is important to track all the spatial transformations that may have affected the accuracy of the parameter. If we use spatial interpolation to increase the resolution of data, we introduce a certain estimation error.

5.3.2 Temporal data analytics

GPS consists of a system of satellites installed by the US Department of Defense [124]. It provides location and time information for GPS receivers located on the Earth. To use this service, devices such as traveling wave recorders and lightning sensors are equipped with GPS receivers that supply information about longitude, latitude, and altitude, as well as a precise time tag and GPS clock. All data must be time referenced in a unique fashion [19]. The following factors are important for time correlation of data:

- Time scales: Data can be collected with different time resolution: yearly, monthly, daily, hourly, once every few minutes or even seconds. In addition, different applications may require different rates of data acquisition.
- Time standard: Different data sources use different time standards [125], such as UTC – Coordinated Universal Time, GPS Satellite Time, which use as a reference TAI – International Atomic Time. All time calculations have to be set into a unified time frame.
- Time synchronization protocols: The accuracy of a time stamp is highly dependent on the type of the signal that is used for time synchronization. Different measuring devices that use GPS synchronization can use different synchronization signals, such as NTP – Network Time Protocol [126] or PTP – Precision Time Protocol [127].

Time scales for different applications and events of interest are presented in Fig. 13, [128]. The temporal correlation module ensures that all the data has a unique temporal reference and that the same time standards and formats are followed. It contains two major parts: 1) historical data processing, and 2) real-time data processing. The final product of historical data processing is a

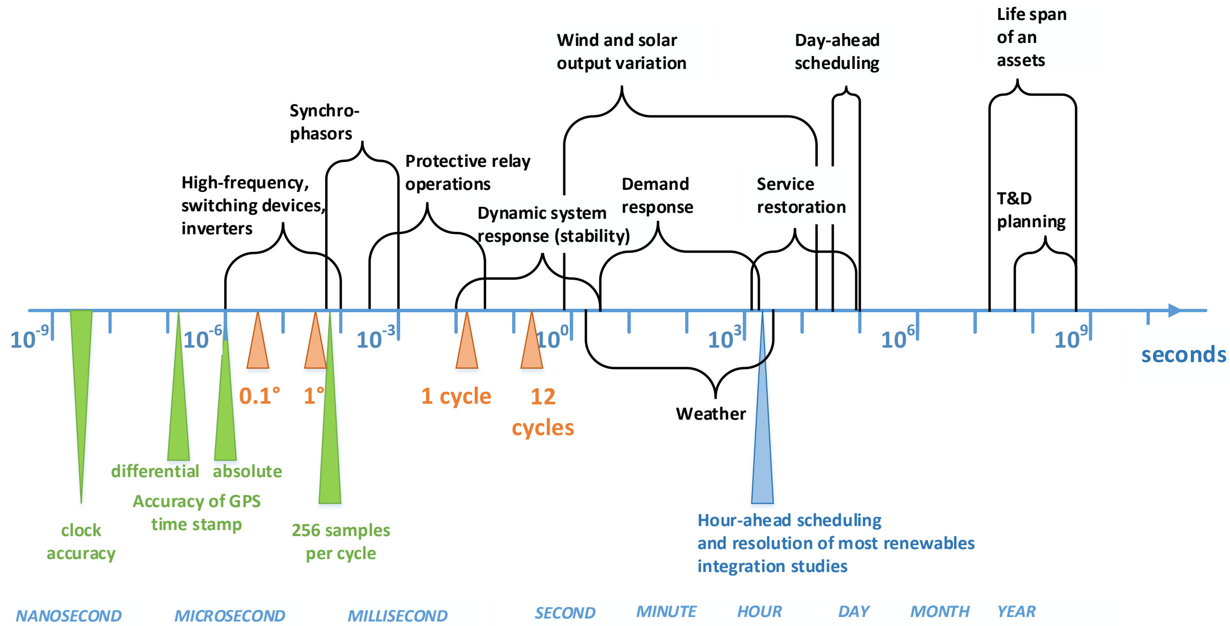

Figure 13 Time scales of the Big Data and applications of interest to the power sector, adapted from [128]

training list for the prediction algorithm. The prediction algorithm uses the time reference to iterate over the historical events, and set the prediction time horizon. The real-time data processing generates input data for the real-time risk maps. The temporal resolution is guided by the availability of different datasets. Similar to the spatial correlation of data, in some cases measurements will be taken to increase or decrease the temporal resolution of datasets by different methods of interpolation (if we want to increase the resolution) or different methods of averaging (if we want to decrease the resolution). The temporal correlation module makes sure that all the data used in the prediction algorithm is set to the same time frame, leading the prediction from one event to another with every parameter being projected to the exact time of the event.

5.4 Conclusion

Based on the discussions in this chapter, it becomes clear why Big Data processing is needed, why it is beneficial to correlate the data, and how data analytics benefits from this approach. The key findings are:

- With Big Data it is possible to collect all the available measurements coming from variety of sources. We develop a unified data model that collects, preprocesses, and spatiotemporally correlates a variety of different data sources for multiple applications. This enables an increase of the overall value of the data for utilities.
- Big Data can handle the variety, volume, and velocity of all the datasets. Weather data sets contain multiple terabytes of data, as well as high resolution imagery files for the extraction of vegetation indices. The measurements are collected on multiple temporal scales, and the solution is capable of processing this in real-time.
- Spatiotemporal correlation of data enables a systematic approach in the creation of necessary inputs for the prediction algorithm. This part of processing makes sure that the prediction algorithm gets a full set of variables that are all associated with a specific location and time of events of interest.

The result of the data management level is a complete set of spatiotemporally correlated parameters that are used as inputs to the prediction algorithm that will be described in Chapter VI, and used in Chapters VIII and IX for two different applications.

CHAPTER VI

DATA ANALYTICS*

6.1 Introduction

In this chapter we describe the prediction algorithm used for this study. The prediction model used for this study is GCRF [129], and it is a structured regression prediction algorithm that incorporates the spatial dependencies between nodes. The benefits of this algorithm for this particular set of applications in power systems are:

- *Graph structure* of structured regression algorithm fits into the electric network model.
- *Spatial interdependencies* of CRF enable extraction of additional knowledge from the data which improves accuracy and provides robustness to missing and bad data.
- *Fast execution* is achieved by constructing feature functions as quadratic functions of the output, making it a Gaussian CRF.

The purpose of the predictive risk model is to estimate what is the probability of a specific event (outage, asset failure) in the network at the specific location for a specific moment in time. The algorithm does this by learning from an extensive set of historical data that includes measured and spatiotemporally correlated parameters for historical events in the network such as outages. Based on this predicted probability, we generate risk maps that represent the state of the network and each individual component for a certain moment in time. Risk maps [130] provide the expected probability of an event (e.g., weather-caused outage) for each component in the network. Risk

---

* This section is in part a reprint with permission of the material in the following papers: (1) T. Dokic, M. Kezunovic, "Predictive Risk Management for Dynamic Tree Trimming Scheduling for Distribution Networks," IEEE Transactions on Smart Grid, September 2018. Copyright 2018, IEEE; (2) M. Kezunovic, Z. Obradovic, T. Dokic, B. Zhang, J. Stojanovic, P. Dehghanian, and P. -C. Chen, "Predicting Spatiotemporal Impacts of Weather on Power Systems using Big Data Science," Pedrycz, Witold, Chen, Shyi-Ming (Eds.), Springer Verlag, Data Science and Big Data: An Environment of Computational Intelligence, ISBN 978-3-319-53474-9, 2017.

mapping is used in many engineering disciplines as a planning and design tool for increasing the robustness of critical infrastructures. Such risk maps can be used by operators in real-time to assess the current state of the network and expected changes in the following several time steps (next hour, next 3 hours, next day). The risk maps are created dynamically as new weather forecasts become available.

6.2 Predictive Risk Framework

Fig. 14 presents an overview of the predictive spatiotemporal risk model [53]. For every moment of time, each network component is assigned a state of risk value. To enable spatiotemporal analysis, the state of risk R is defined as follows [131]:

$$R(G,t) = P[T(G,t)] \cdot P[C(G,t)|T(G,t)] \tag{7}$$

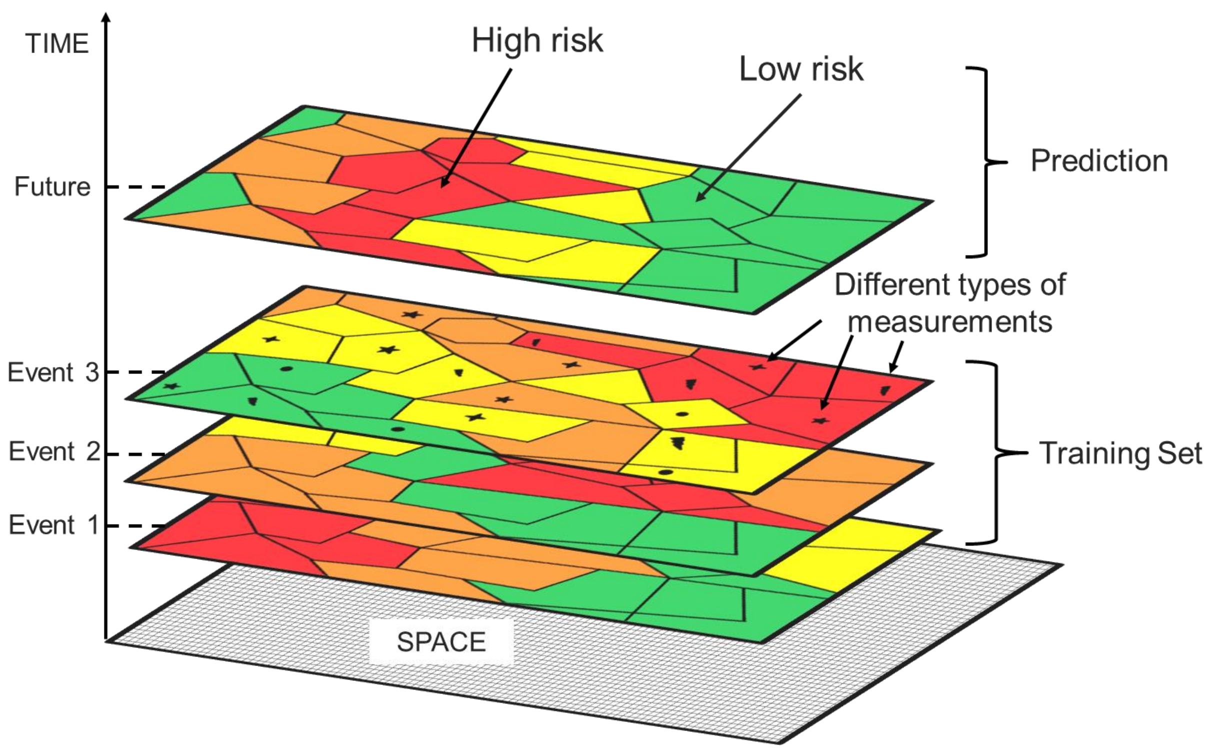

Figure 14 Spatiotemporal Prediction Model, reprinted from [5]

where G represents the longitude and latitude of a single element, and t represents the moment in time for which the observation is made. A unique state of risk value is assigned to each distribution feeder section. $T(G,t)$ represents the threat intensity. Threat intensity is defined as a qualitative metric of the weather condition severity. The first term in (7), $P[T(G,t)]$, is a hazard probability. This term represents the probability of a severe weather condition with the selected threat intensity. The second term, $P[C(G,t)|T(G,t)]$, is network vulnerability, where $C(G,t)$ is an occurrence of a consequence. Vulnerability is a conditional probability of a consequence (such as a vegetation-caused outage) in the distribution network if and when severe weather is present. The risk definition presented here is an adaptation of definition in [132], where the last part of the risk-economic impact is not included. In this research, the economic impact is calculated separately and included in the optimal maintenance model as one of constraints. The economic impact calculation takes the predicted risk values as inputs to the optimization objective function. The optimization will use the calculated risk maps for a period of time to determine the best set of countermeasures that can improve the reliability of the system. The calculated risk maps are used as inputs to the optimization objective function that is trying to minimize the overall risk for the network. The details of how the economic impact is combined with the risk framework will be described in the Chapter VII.

6.3 Prediction Model

6.3.1 Continuous Conditional Random Fields

As we discussed in Chapter 2.6, the problem at hand can be modeled using graph-based structured regression. Since we want to use interdependencies between nodes to improve prediction accuracy and enable robust prediction in case of bad and missing data, we select the Conditional Random Fields as the algorithm. Continuous Conditional Random Fields, as presented in Fig. 15 [133], model the conditional distribution $P(\boldsymbol{y}|X)$ over a set of outputs $\boldsymbol{y}$ given all inputs X, as

$$P(\boldsymbol{y}|X) = \frac{1}{Z(X, \boldsymbol{\alpha}, \boldsymbol{\beta})} \exp\big(\phi(\boldsymbol{y}, X, \boldsymbol{\alpha}, \boldsymbol{\beta})\big) \tag{8}$$

Where the term in the exponent $\phi(\boldsymbol{y}, X, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is defined as

$$\phi(\boldsymbol{y}, X, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{N} A(\boldsymbol{\alpha}, y_i, X) + \sum_{j \sim i} I\big(\boldsymbol{\beta}, y_i, y_j, x\big) \tag{9}$$

and the normalization constant $Z(X, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is defined as

$$Z(X, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int_{\boldsymbol{y}} \exp\big(\phi(\boldsymbol{y}, X, \boldsymbol{\alpha}, \boldsymbol{\beta})\big) d\boldsymbol{y} \tag{10}$$

$A(\boldsymbol{\alpha}, y_i, X)$ is a real-valued function called the association potential, and models how output y_i is associated with the set of inputs X, where $\boldsymbol{\alpha}$ is an K-dimensional set of parameters [131].

$I\big(\boldsymbol{\beta}, y_i, y_j, X\big)$ is a real-valued function called the interaction potential that describes the interactions between pairs of outputs $y_i \sim y_j$, where $\boldsymbol{\beta}$ is an L-dimensional set of parameters [131]. This parameter is of a great importance for our studies, since the electric network is represented as the connected graph where every event happening in one node has impact on all the neighborhood nodes.

Learning is accomplished by finding values of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for which the conditional log-likelihood of the set of training examples is maximized, as in Eqs. 11 and 12:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum \log P(\boldsymbol{y}|X) \tag{11}$$

$$\big(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}\big) = \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \big(L(\boldsymbol{\alpha}, \boldsymbol{\beta})\big) \tag{12}$$

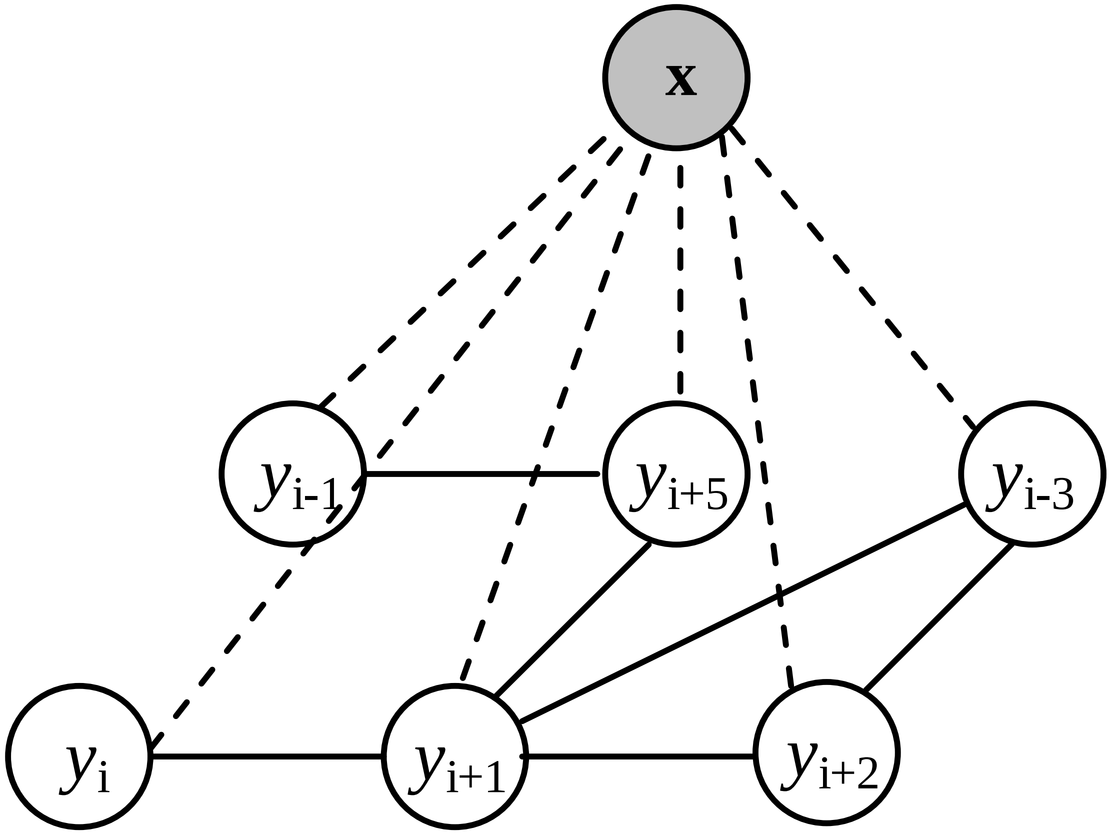

Figure 15 Continuous CRF graphical structure, reprinted from [131]

The values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be chosen using optimization algorithms, for example gradient ascent optimization. The inference task is to find the outputs $\boldsymbol{y}$ for a given set of observations X and estimated parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that the conditional probability $P(y|X)$ is maximized as in Eq. 13 [131]:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|X) \tag{13}$$

In CCRF applications, the parameters $A(\boldsymbol{\alpha}, y_i, X)$ and $I\left(\boldsymbol{\beta}, y_i, y_j, X\right)$ are often defined as linear combinations of a set of feature functions f and g in terms of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ [131]:

$$A(\boldsymbol{\alpha}, y_i, X) = \sum_{k=1}^{K} \boldsymbol{\alpha}_k\, f_k(y_i, X) \tag{14}$$

$$I(\boldsymbol{\beta}, y_i, y_j, x) = \sum_{l=1}^{L} \boldsymbol{\beta}_l\, g_l(y_i, y_j, X) \tag{15}$$

When using feature functions, any potentially relevant feature could be included in the model because parameter estimation automatically determines their actual relevance by feature weighting. This allows modeling of arbitrary relationships between inputs and outputs [133].

6.3.2 Gaussian Conditional Random Fields

The applications of interest for this study include processing on large electric networks where every component (tower, pole, one line span) is modeled as one node of the graph, creating large graphs with hundreds of thousands of nodes, that are observed in tens of thousandths of time instances for a variety of parameters. The construction of feature functions A and I in CRF can be done many different ways; however, to reduce the computational complexity of learning and inference, A and I can be constructed as quadratic functions of y [129]. If A and I are quadratic functions of y, $P(y|X)$ becomes a Gaussian distribution – leading to the name Gaussian Conditional Random Fields (GCRF), Fig. 16. This improves the speed of execution of the program significantly, making it possible to process the data in real-time for the entire network.

Let us assume we are given K unstructured predictors, $R_k(X), k = 1, \dots K$, that predict a single output y_i taking into account X. To model the dependency between the prediction and output, quadratic feature functions are introduced [129]:

$$f_k(y_i,) = -\left(y_i - R_k(X)\right)^2, k = 1, \dots K \tag{16}$$

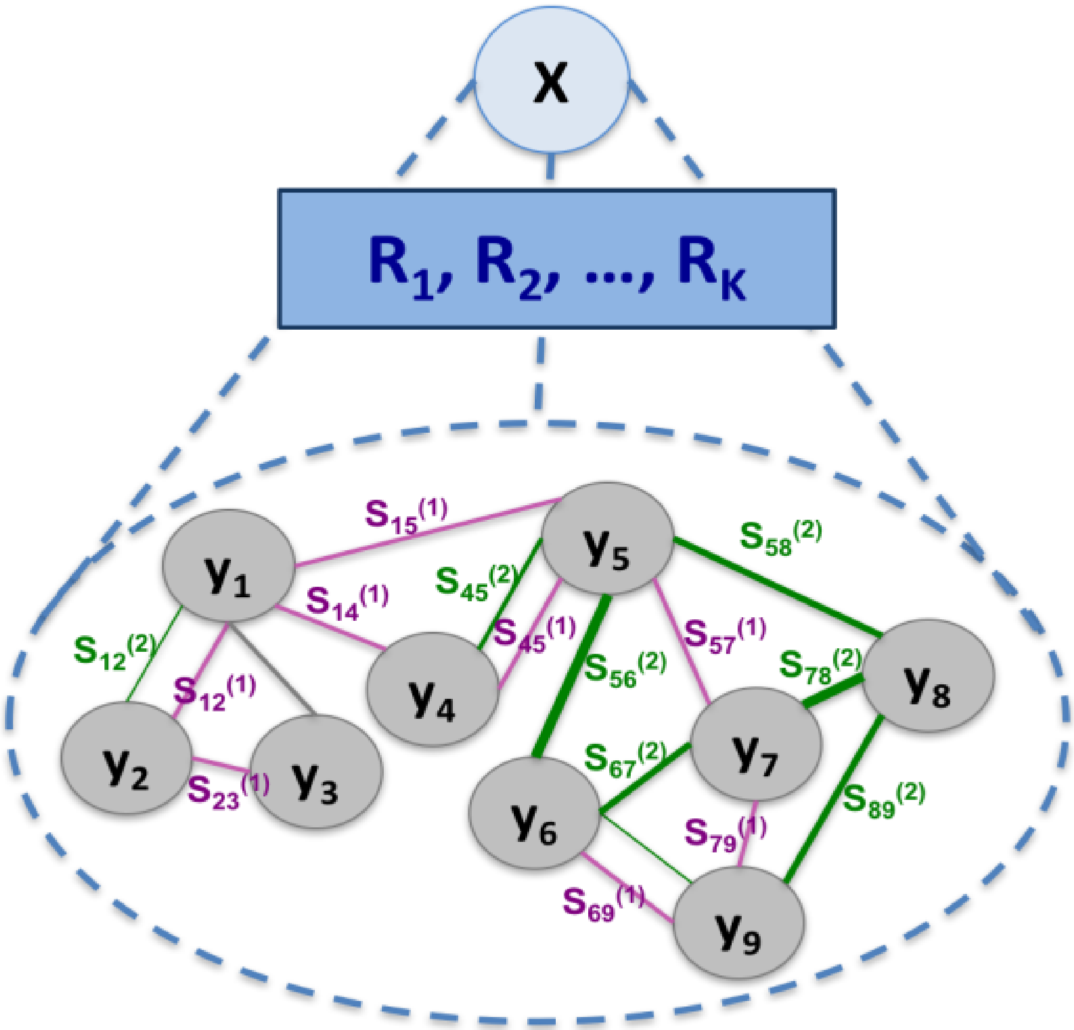

Figure 16 Gaussian CRF graphical structure, reprinted from [131]

These feature functions describe the association potentials. Their values are large when predictions and outputs are similar. The association potential will identify the association between every independent variable and the output. For example, it models the association between measured temperature and insulator strength. To model the correlation among outputs, quadratic feature functions are introduced [131]:

$$g_l(y_i, y_j, X) = -e_{ij}^{(l)} S_{ij}^{(l)}(X)\left(y_i - y_j\right)^2, \;\; e_{ij}^{(l)} = 1, \text{ if } (i,j) \in G_l, \quad e_{ij}^{(l)} = 0, \text{ otherwise} \tag{17}$$

This imposes that outputs y_i and y_j have similar values if they are connected with a link of a graph. $S_{ij}^{(l)}(X)$ represents the similarity between outputs y_i and y_j. The larger $S_{ij}^{(l)}(X)$ is, the greater similarity between the outputs y_i and y_j.

$P(\boldsymbol{y}|X)$ for the CRF model (4.2), which uses quadratic feature functions, can be represented as a multivariate Gaussian distribution. The resulting CRF model can be written as in [131]:

$$P(y|X) = \frac{1}{Z}\exp(-\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_k\left(y_i - R_k(X)\right)^2 - \sum_{i,j}\sum_{l=1}^{L}\beta_l e_{ij}^{(l)} S_{ij}^{(l)}(X)\left(y_i - y_j\right)^2) \tag{18}$$

The learning task is to choose $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to maximize the conditional log-likelihood,

$$(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\arg\max}\left(L(\boldsymbol{\alpha}, \boldsymbol{\beta})\right) \text{ where } L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum \log P(\boldsymbol{y}|X) \tag{19}$$

For the model to be feasible, we can impose the constraint that all elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are greater than 0, which results in a constrained optimization problem. To convert it to an unconstrained optimization problem, a technique used in [135] is adapted. Then, all parameters are learned by the gradient-based optimization. [135].

6.4 Conclusion

The presented prediction model performs well with the application of interest to asset and outage management in power systems due to the following reasons: 1) it is capable of following the temporal progression of events; 2) the data structure is a weighted connected graph which perfectly represents the electric network and models all the spatial dependencies in it; 3) events in the power system and their impact often impact a number of neighboring nodes. For example, weather-related outages may affect a certain area, or lightning backflashover may propagate through the network from its source until it attenuates.

The results of the prediction algorithm are used to generate various event risk maps which provide operators with valuable information about what may happen in the network in the near future. However, it is of most importance to also know what kinds of actions can be taken to improve the future performance of the network. With this purpose, the risk maps are observed over time to generate optimal action schedules for smart asset management, which will be described in the next Chapter VII. This gives the main purpose of the predictive risk analysis, where the optimal actions are chosen with a goal of reducing the future risk in the network.

CHAPTER VII

ECONOMIC ASSESSMENT*

7.1 Introduction

The predicted levels of risk over the network for a specified period of time (fiscal quarter, 6 months, 1 year, etc.) can be used to develop optimal asset management strategies for different components in the network. This feature makes predictive risk maps of great benefit, since they do not just predict the expected levels of risk in the network; they can be used to automatically decide the appropriate countermeasures that will reduce the risk in the future. For example, in operations, risk maps can be used to assess the state of the network, providing operators with better situational awareness. Operators can make better decisions about network configuration changes and dispatch of maintenance trucks. Asset management groups can use the risk maps to analyze the changes in states of various assets and develop techniques for optimal maintenance and replacement. Customers can use the risk maps to analyze the expected probability of losing service in near future, which can help in making economically efficient decisions for their business. The predictive risk maps can be generated with multiple temporal scales, and thus are able to serve many different applications in the electric utility, as demonstrated in Fig. 13 in Chapter 5.3.2. A maintenance scheduler has a goal to minimize the risk for the whole network while spending only the predetermined maintenance budget. While the objective and constraints of the optimization problem can be chosen in many different ways, we chose the approach to minimize risk with economic constraints because it is the most common approach taken by the utilities that are

* This section is in part a reprint with permission of the material in the following paper: T. Dokic, M. Kezunovic, "Optimized Asset Management in Distribution Systems Based on Predictive Risk Analysis," Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion - MEDPOWER, Dubrovnik, Croatia, November 2018.

deciding the budget for each department in advance on a yearly basis. Two types of maintenance costs are identified: 1) planned maintenance typically has a predetermined budget, and is performed periodically; 2) reactive maintenance includes the actions that occurred after the unexpected outage or asset failure, and the budget is variable. The optimal maintenance scheduler that will be described in this chapter aims at staying within the limits of the planned maintenance budget, while reducing the reactive maintenance costs.

7.2 Optimization Framework

The specific optimization problem has to be defined separately for each distribution asset type (pole, transformer, insulator…) but the overall procedure can be defined as follows: minimize the total risk for the network [99]:

$$\mathrm{MIN}\{\mathrm{SUM}(RISK_{\mathrm{COMPONENT,\ TIME}})\} \tag{20}$$

subject to following economic constraint:

$$\{\mathrm{SUM}(ACTION\_COST)\} \leq BUDGET \tag{21}$$

The optimization problem solver will iterate over various actions (component maintenance, component repair, component replacement, environment assessment such as tree trimming, etc.) until it finds the best asset management schedule. For each time step, each component has an action flag that indicates if there is an action on that component and what type of action is performed. This makes the optimization problem nonlinear. To provide a feasible solution in time, heuristic solvers need to be considered.

The impacts of the environment and component vulnerabilities for each moment in time are accumulated in the dynamic risk value we are trying to minimize. The limits of the budget for periodic (planned) actions are taken into account as constraints.

In addition to the main objective – minimizing the risk – reactive maintenance is indirectly targeted for minimization, and it is used for additional validation and testing of this approach's performance. Our goal is, by minimizing the overall network risk, to also minimize the cost of reactive maintenance. As part of validation, after the optimization problem is solved, we compare the reactive asset management cost that was spent during the period of interest to the evaluated reactive maintenance expense that would be spent if an optimal asset management schedule was followed.

Following are the required steps of the optimal maintenance scheduler, as presented in Fig. 17:

1. Generate risk maps based on the historical data and weather forecast and store the risk value for each component in the network. This step contains three tasks:
   a) Calculate the weather hazard using the weather forecast [32]. In this step we are evaluating the expected unfolding weather conditions that will affect the network at a certain moment in time.
   b) Calculate network vulnerability using historical data and the current profile of the network and environment. In this step we learn from the historical outage and weather data [30, 31] what the vulnerabilities of the network are, and we calculate, based on the knowledge from the past, what is the probability of an outage under an unfolding weather condition.
   c) Generate action on a specific component. By performing any of the countermeasures, it is possible to reduce the network vulnerability to the unfolding weather conditions. The optimization algorithm will iterate over multiple action configurations until it finds the optimal schedule.

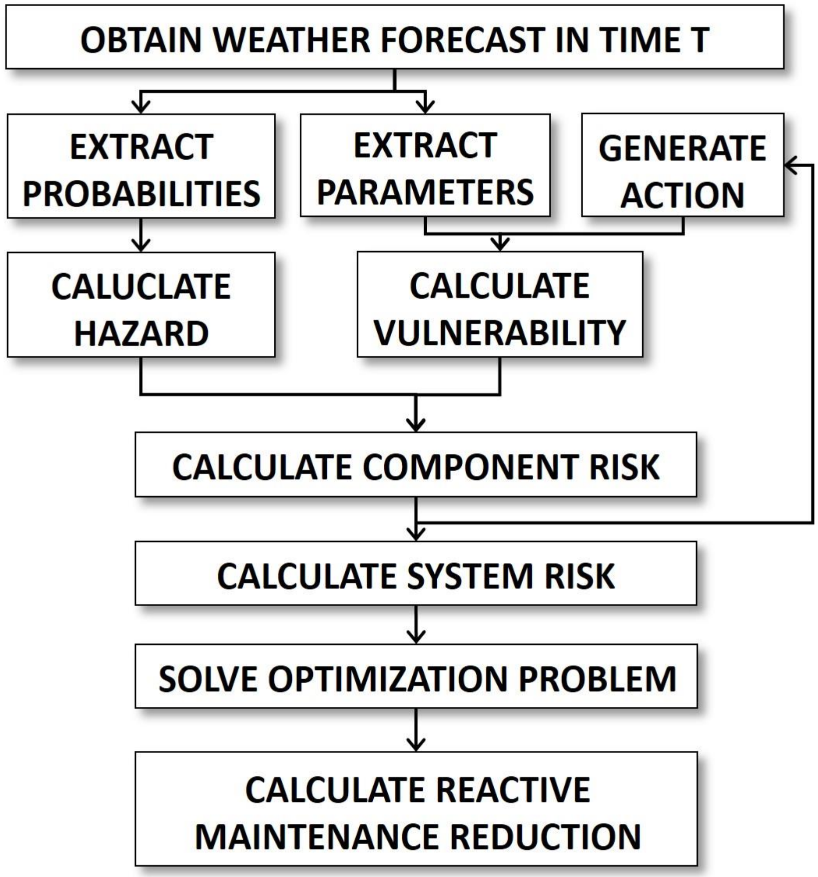

Figure 17 Optimal Risk-based Scheduler, reprinted from [99]

2. Calculate the system risk by averaging or summarizing the risk over all components.
3. Define the optimization problem that minimizes the calculated system-level risk. In this step, appropriate countermeasures need to be selected. For example, if we are observing vegetation management, the main countermeasure would be tree trimming. In another example, if we are targeting insulators, countermeasures may include insulator cleaning, insulator repair, insulator replacement, etc.

4. Set the optimization constraints to limit the periodic asset management expense. In this step the specific practices followed by the utility need to be observed to set realistic economic constraints.
5. Solve the nonlinear optimization problem by applying heuristics (Lagrangian Relaxation, Support Vector Machines, Neural Networks, etc.). In this dissertation the optimization problem is solved using enhanced linear programming relaxation with Lagrangian relaxation, with the addition of the heuristic method described in detail in [136]. This is a common approach for nonlinear problems that have risk reduction as the optimization objective, and economic impacts embedded into constraints.
6. Calculate the reduction in reactive maintenance cost after the outage. During the validation process, the reduction in reactive maintenance can only be estimated. After the deployment in the field, the testing process can observe the changes in reactive maintenance expense before and after dynamic maintenance scheduling.

7.3 Conclusion

Optimal asset management based on predictive risk analysis provides a new paradigm that generates optimal maintenance tasks based on predicted levels of risk in the network in the future. The described optimal asset management system from this research can be used in different applications with different optimization targets and constraints, as long as the objective is to minimize the risk of a specific event, and we need to stay within a predetermined maintenance budget.

The specific definition of risk needs to be created for each case individually. In the following chapters we will demonstrate the definition of the optimization problem for three different applications: 1) optimal tree trimming scheduling, 2) optimal maintenance/replacement of insulators, and 3) optimal placement of LSAs. With the description of optimization process in this chapter we conclude the three-level methodology. In Chapters VIII and IX we will demonstrate how all three levels can be implemented for two different applications: 1) vegetation management, and 2) insulation coordination.

CHAPTER VIII

APPLICATION TO VEGETATION MANAGEMENT*

8.1 Introduction

This chapter demonstrates the application of the developed three-level Big Data analytics framework to predictive vegetation management in distribution. This demonstration serves as a validation of our hypothesis that more accurate predictions are possible by structured learning from merged heterogeneous Big Data. This will be demonstrated by evaluating the prediction accuracy of the GCRF algorithm for vegetation-related faults, and the expected improvements in network reliability after an optimal tree trimming schedule is applied.

The application discussed in this chapter differentiates itself from other vegetation management solutions by the use of an extensive set of data and specifically tailored data analytics to create predictions. We correlate different datasets and use them as inputs to the new predictive risk model that utilizes spatiotemporal data to produce real-time risk maps for tree trimming in the distribution network. The prediction algorithm, based on a GCRF model described in Chapter VI, leverages the spatial similarities between different feeder sections to ensure better prediction performance and compensate for missing data. The resulting risk model allows the implementation of a dynamically changing trimming scheduler that optimizes the tree trimming process. It will be shown that the achieved reduction in risk has the potential of reducing the cost of reactive tree trimming. The method is applied to a real distribution network and utility data. The testing

---

* This section is in part a reprint with permission of the material in the following paper: T. Dokic, M. Kezunovic, "Predictive Risk Management for Dynamic Tree Trimming Scheduling for Distribution Networks," IEEE Transactions on Smart Grid, September 2018. Copyright 2018, IEEE.

confirms that the outages occurred in zones with risk predicted to be greater than 64%, which suggests a new predictive paradigm for vegetation management strategies.

8.2 Data

Raw data are processed to remove unused components. All the data that has a geographical reference is placed into a geodatabase during the preprocessing. Table 6 lists the extracted parameters needed for the prediction model, and the associated temporal and spatial references.

Data come with different spatial and temporal resolutions. Historical weather data from ASOS [29] land stations has the highest temporal resolution (up to 1 min); however, the spatial resolution of data is low, including only a few weather stations in the network service area. Vegetation data has a low temporal resolution (collected once per year or two years) but has a high spatial resolution (up to 50 cm). The rate of data collection varies not only between different data sets, but also it can vary within a single data set. For example, weather data is collected by land-based weather stations with a maximum rate of one data point per minute, but the rate can go down

Table 6 Parameters Extracted in Preprocessing, reprinted from [53]

	Historical Outage Data	**Periodic Tree Trimming**	**Reactive Tree Trimming**	**Poles**	**Lines**	**Vegetation**	**Weather**
Spatial	Point shapefile	Polyline shapefile	Polyline shapefile	Point shapefile	Polyline shapefile	• Raster • Polygon shapefiles	• Points • Polygon shapefiles
Temporal	Start and end time	Year quarter	Date	Static	Static	Year	1 min to 3 hours
Other parameters	• Num. of customers • Cause code	• Trim period • Num. of customers • Cost	• Cost	• Material/class • Height	• Conductor size • Conductor count • Conductor material • Nominal voltage	• Imagery • Vegetation classes	• Wind (speed, gust, direction) • Temperature • Precipitation • Humidity • Pressure • Forecast indices

to one data point per hour. In some rare cases, the rate may go as low as one measurement within several hours. After preprocessing, the dataset is still not ready for the input into the predictive risk model. All the parameters need to be spatially and temporally correlated, as is described in 8.2.1 and 8.2.2.

Image data is used to extract the location of vegetation surrounding the network. The imagery is collected from the TNRIS database [35]. The following orthoimagery datasets are used in the study:

- National Agriculture Imagery Program 1m NC\CIR for years 2010, 2012, 2014, and 2016;
- Texas Orthoimagery Program 50cm NC\CIR for 2015.

The datasets are loaded into the geodatabase as raster files. First, to reduce the amount of data for processing, imagery raster files are clipped to a 20 m buffer around the distribution lines. Then unsupervised image classification [137] is applied. The iso-cluster is set to 40 classes in all datasets. In the next step the classes are reclassified to "vegetation class" and "non-vegetation class", and converted into a polygon shapefile. The vegetation class is transferred to the next step

a) b) c)

Figure 18 Example of vegetation extraction: a) 40 classes, b) imagery for reference, and c) reclassified, reprinted from [53] (spatial correlation), and the rest is discarded. In Fig. 18, we provide examples of the unsupervised classification (a), and the resulting map after reclassifying (c). Map (b) is in Fig. 18 for visual reference.

The result of image processing is a set of historical maps with vegetation locations. These maps are then spatially joined with the EMST developed by the TPWD [33]. The EMST data contains classification by vegetation type into 398 distinct classes, out of which 49 classes are present at the network location of interest. The average canopy height for 49 vegetation classes in the network area is then added to the vegetation dataset as a parameter.

8.2.1 Spatial Correlation of Data

The purpose of the spatial correlation module is to provide spatial links between different data sets. For example, for every historical outage we want to know the weather conditions at that specific location, the distance between the line and the closest tree, the location of areas that were trimmed, etc. The spatial correlation module is presented in Fig. 19. We distinguish three parts of the spatial processing module:

Weather data processing encompasses creating the weather data grid that is overlaid on the utility network and has a spatial resolution of 1 km. The weather parameters in each grid cell are calculated from the weather station values using linear interpolation.

Vegetation data processing extracts the vegetation indices, such as distance between the lines and vegetation and growth rate, using spatial links between multiple preprocessed vegetation files. All the calculated parameters are stored as attributes in the final vegetation polyline shapefile.

Utility data processing converts the historical tables to shapefiles identifying the locations of points and polylines based on the line section codes and/or addresses provided in the utility's

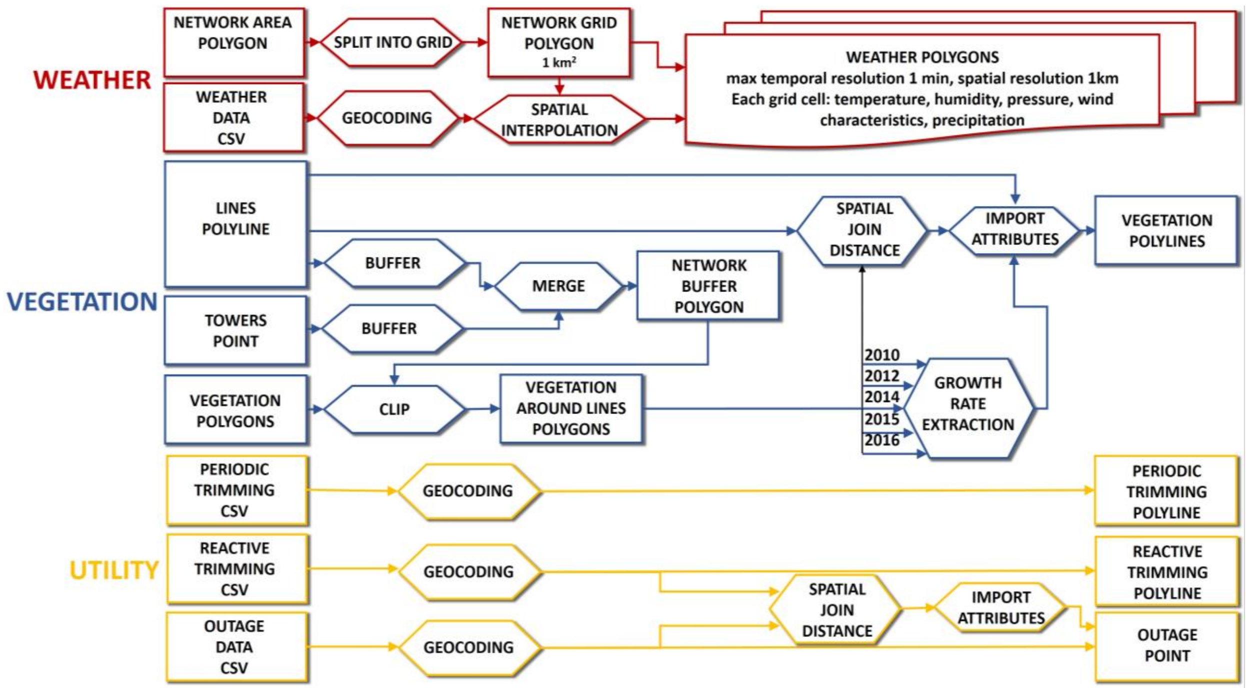

Figure 19 Spatial correlation of data, reprinted from [53]

CSV files. In addition, every reactive tree trimming action is correlated with the outage that lead to it.

To deal with different spatial resolutions of data we used multiple approaches, all included in Fig. 3. We used spatial interpolation where weather data was extracted for every location in the network based on the original weather station data with sparse locations. In other instances, data was projected to a nearby location using a spatial join. For example, the distance between the line and vegetation is projected to the line using a spatial join based on distance.

8.2.2 Temporal Correlation of Data

The temporal correlation module has five historical input datasets (weather, vegetation, outage, periodic tree trimming, and reactive tree trimming), and real-time weather forecast input.

Each dataset contains a variety of parameters (attributes) from Table 6, and is stored as a GIS shapefile. Static datasets (network feeders and poles) are assumed not to change over the observed period, and do not require any temporal correlation. Fig. 20 presents an overview of the temporal correlation module containing two major parts: 1) historical data processing, and 2) real-time data processing. The final product of historical data processing is a training list for the prediction algorithm. The real-time data processing generates input data for the real-time risk maps by generating the data for hazard, vulnerability, and economic impact that feeds the dynamic tree trimming scheduler, which will be described in the following sections.

The temporal resolution is guided by the occurrence of outages. For every outage we want to extract, all the relevant information is included as presented in Fig. 20. For each outage, the data points that are closest in time are chosen from each set individually. For example, in the case of historical weather data, the closest data points were within one minute of outage. On the other hand, the closest vegetation maps could be up to several weeks apart.

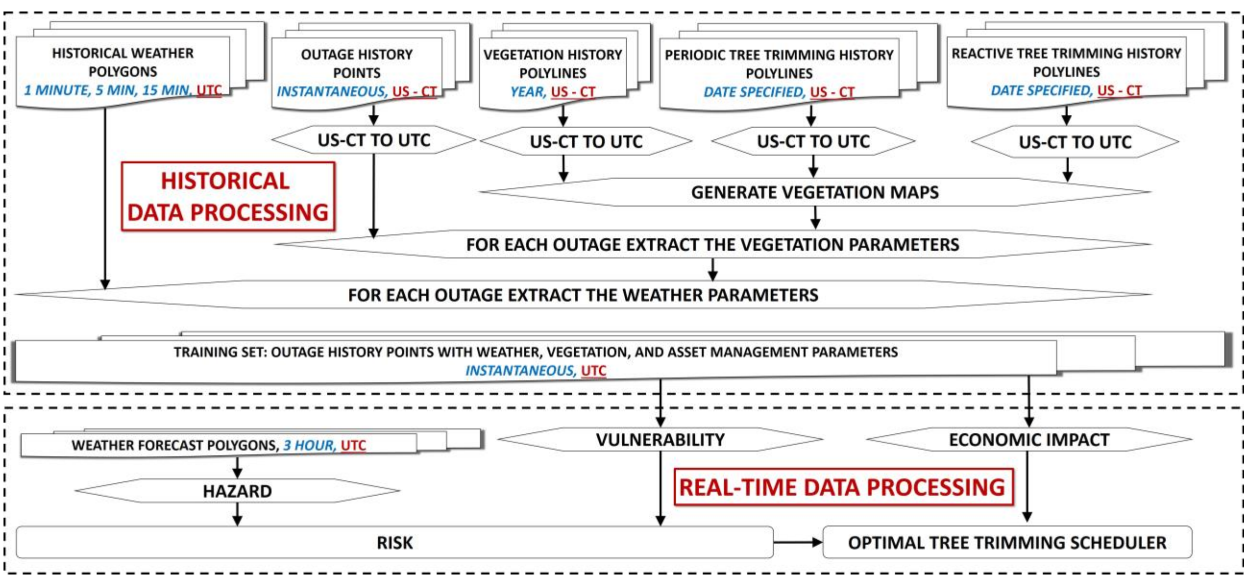

Figure 20 Temporal correlation of data, reprinted from [53]

8.3 Risk for Vegetation Management

8.3.1 Hazard

The hazard level is calculated based on the weather forecast data for a specific time and location. The data from the NDFD [32] is used. The database contains the forecast up to 7 days in the future with a time resolution of 3 hours. The updated forecast is provided every 3 hours. The spatial resolution of the weather forecast data is 5 km. Because the weather forecast data is updated every 3 hours with a maximum resolution of 3 hours, the risk maps are generated with the same 3 hours resolution.

Figures 7-9 summarize the construction of weather hazard. The following parameters are observed: wind speed, direction, and gust, temperature, relative humidity, convective hazard outlook, probability of critical fire, probability of dry lightning, hail probability, tornado probability, probability of severe thunderstorms, damaging thunderstorm wind probability, and extreme hail probability. Based on the values of the observed parameters, the threat level is classified into 6 groups from 0 to 5, where 0 represents normal weather conditions without any potentially severe elements, and 5 represents extremely severe weather conditions. k-means clustering [138] was used for classification into 6 groups. k-means clustering enables the

Table 7 Hazard Classification, reprinted from [53]

Probability [%]	**Threat Intensity**					
	0	**1**	**2**	**3**	**4**	**5**
0-20						
20-40						
40-60						
60-80						
80-100						

construction of hazard consequence levels from the individual weather parameters. This way, multiple parameters are combined into a single parameter, *Threat Intensity*, with 6 different states. The clustering is done using historical weather data, where different configurations of weather parameters are associated with their measured impact on the outage occurrence. Then the Hazard is constructed as a heat map in Table 7, where each threat level has an assigned probability of occurrence determined based on weather forecast. The construction of the heat map is based on [130], where heat maps are constructed following two steps: 1) constructing the probability matrix

Table 8 Probability of Threat Level Occurrence, reprinted from [53]

Probability Range [%]	**Description**
0-20	Extremely Unlikely
20-40	Highly Unlikely
40-60	Doubtful
60-80	Somewhat Likely
80-100	Very Likely

Table 9 Threat Intensity Levels, reprinted from [53]

Category	**Description**	**Example**
0	None	No impact on the network
1	Minor	Minor service interruptions, no restoration needed
2	Moderate	Some outages in the network, some restoration needed
3	Low Severe	Moderate number of outages in the network, restoration delays may occur, e.g. rainy weather
4	High Severe	Multiple outages in the network with longer restoration duration, e.g. thunderstorm
5	Catastrophic	The whole network or very large parts of the network under the disconnected – large blackouts, e.g. Hurricane

as in Table 8, and 2) constructing the threat intensity matrix as in Table 9. The Hazard value ranges from extremely low, marked green in Table 7, to extremely high, marked as red.

8.3.2 Vulnerability

The GCRF predicts the state of vegetation impact, denoted y, based on historical measurements in the input vector *X*. The following historical measurements are stored in the input vector x: wind speed, wind direction, wind gust, precipitation, temperature, humidity, pressure, vegetation distance to the line section, vegetation spread, vegetation growth rate, vegetation health index, pole height, tree trimming period, time since last tree trimming, outage duration, number of customers affected. The output y of the algorithm is the predicted state of vegetation impact on the feeder section.

Historical outages are an integral part of the Vulnerability. The prediction of future vulnerability is done based on the knowledge collected from previous outages. As listed in Table 6, the historical outage data contains information about the duration of the outage and the number of customers affected by it. This information guides the prediction algorithm to generate higher vulnerability levels in the cases where more customers were affected by the outage and for the greater duration.

8.4 Optimal Tree Trimming

The goal of the optimization model is to minimize the overall risk of the outage in the system while maintaining the budget allocated for periodic tree trimming. To achieve that, the quarterly periodic tree trimming schedule is designed based on the risk prediction for the next 3 months. The time instances when the risk map is created are every three hours during a three-month period. A total of T time instances is created each quarter. The risk is calculated for each of the N feeder sections. An optimized tree trimming schedule is determined by solving the following optimization problem:

$$max\left\{R = \sum_{t=1}^{T} \frac{1}{N} \sum_{n=1}^{N} \Delta R_{n,t} \cdot F_{n,t}\right\} \tag{22}$$

$$F_{n,t} = \begin{cases} 0, & \text{section n not trimmed at time t} \\ 1, & \text{section n is trimmed at time t} \end{cases}$$

where $\Delta R_{n,\theta} = R_{n,(\theta-1)} - R_{n,\theta}$ is the difference in risk value for feeder n before and after the tree trimming is performed. The following constraints are enforced:

$$\sum_{t=1}^{T} \sum_{n=1}^{N} F_{n,t} \cdot PC_{n,t} \leq PA \tag{23}$$

$$\text{For t=1,…,T, } \textstyle\sum_{n=1}^{N} F_{n,t} \leq 1 \tag{24}$$

where *R* is a total reduction in risk, $PC_{n,t}$ is the cost of tree trimming of section n at the time instance *t*; and *PA* is a total budget allocated for periodic tree trimming during the observed quarter. The optimization problem is nonlinear, and it is solved using enhanced linear programming relaxation with Lagrangian relaxation plus the heuristic method described in [136].

While a reduction in reactive tree trimming cost is not an explicit goal of the optimization problem, it is still calculated to check the impact of risk reduction on the reactive tree trimming cost. To do that, the reactive tree trimming orders are iterated, and for each one it is checked if the developed tree trimming scheduler recommended trimming of the area prior to the outage. If an area is part of the recommended tree trimming schedule in a time frame before the reactive tree trimming was performed, the reactive tree trimming cost is deducted from the total.

8.5 Evaluation and Results

The observed utility distribution network has an area of ~2,000 km^2. It contains ~200,000 poles, and ~120,000 lines. The historical outage and weather data were collected for the period from January 2011 up to the end of April 2016. Over this period, 505 weather-related outages have been observed in the area, where a total of 331 outages were vegetation-caused (Fig 21). The training set for a prediction algorithm consists of the first 300 historical outages in temporal order.

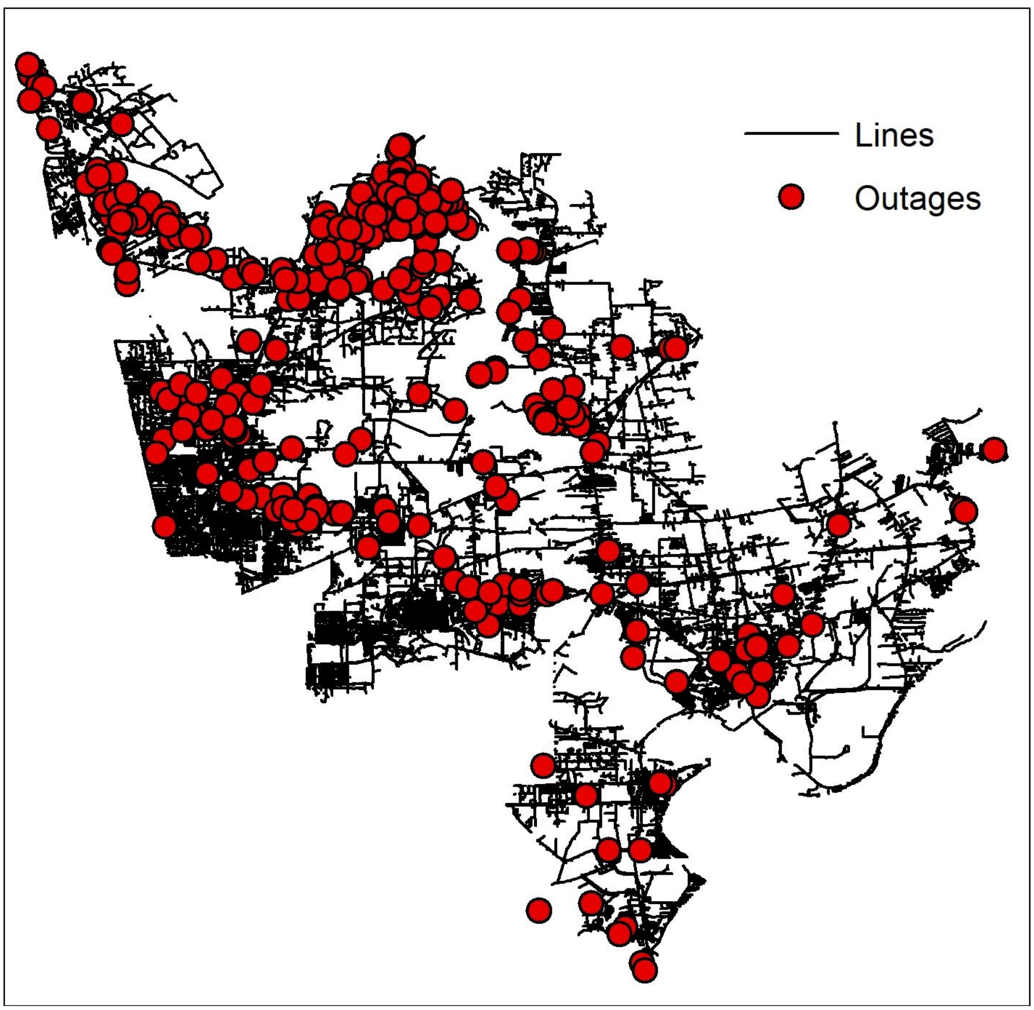

Figure 21 Distribution of historical vegetation outages, reprinted from [53]

The remaining 31 outages that occurred at the end of 2015 and beginning of 2016 are used as testing set.

The example of the predicted Hazard and Vulnerability map for an outage event that occurred on February 23, 2016 is presented in Fig. 22 and Fig. 23 respectively. The weather hazard is presented as a grid covering the area of the network, while the vulnerability is assigned to each line section individually. The resulting predicted risk map for the observed date is presented in

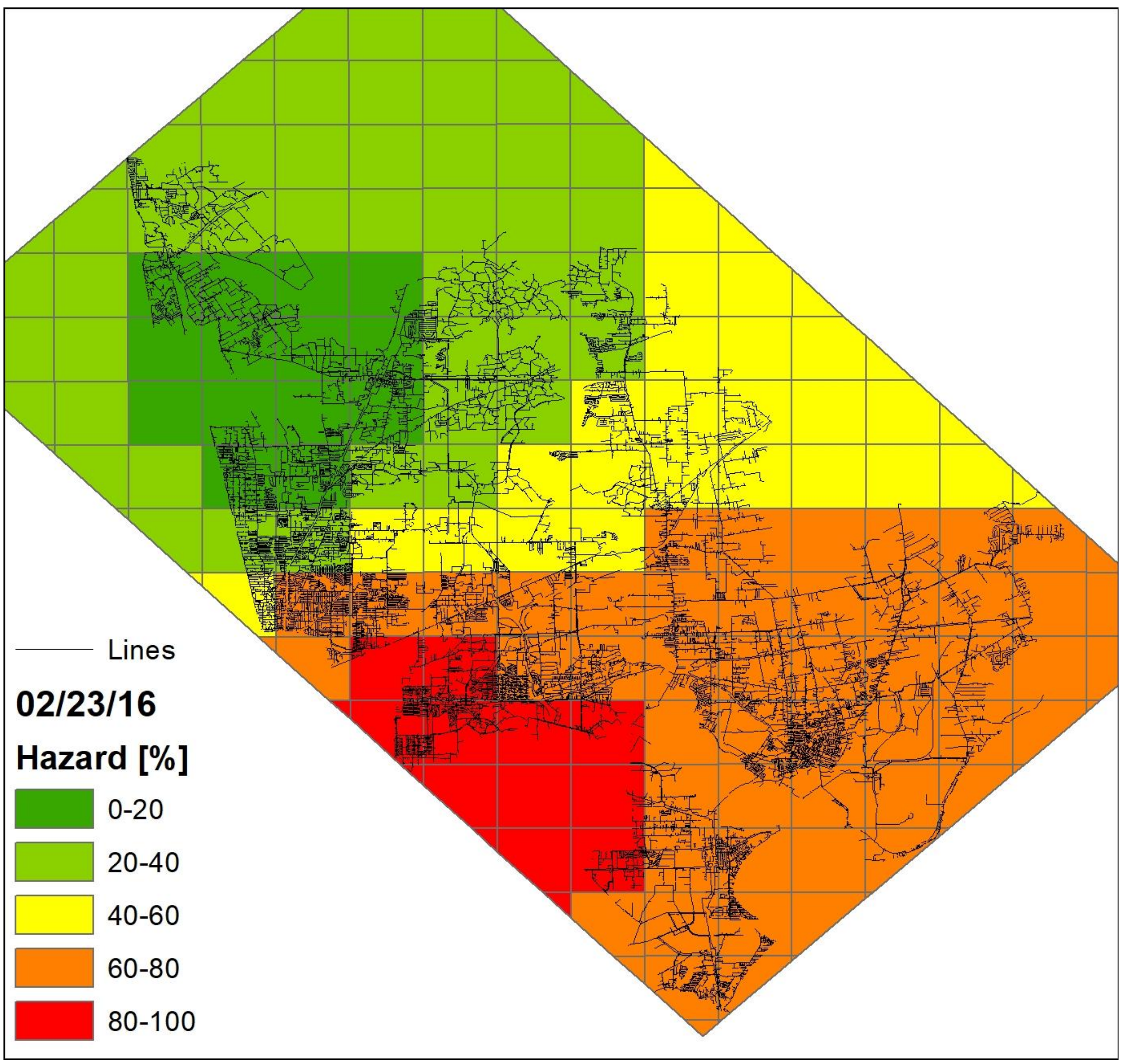

Figure 22 Hazard Map for 02/23/2016, reprinted from [53]

Fig. 24. As it can be seen in the upper right corner the predicted risk value on the faulted section for the outage in Fig. 24 that occurred on 02/23/2016 was 84%.

The predicted risk values for all 31 test outages are presented in Fig. 25. The minimum risk value during an outage is 64%. There are 6 instances for which the risk probability was less than 75%, all of which occur during the days with a low weather hazard. We would like to speculate that in the absence of weather hazard information, when the algorithm is limited to predicting the

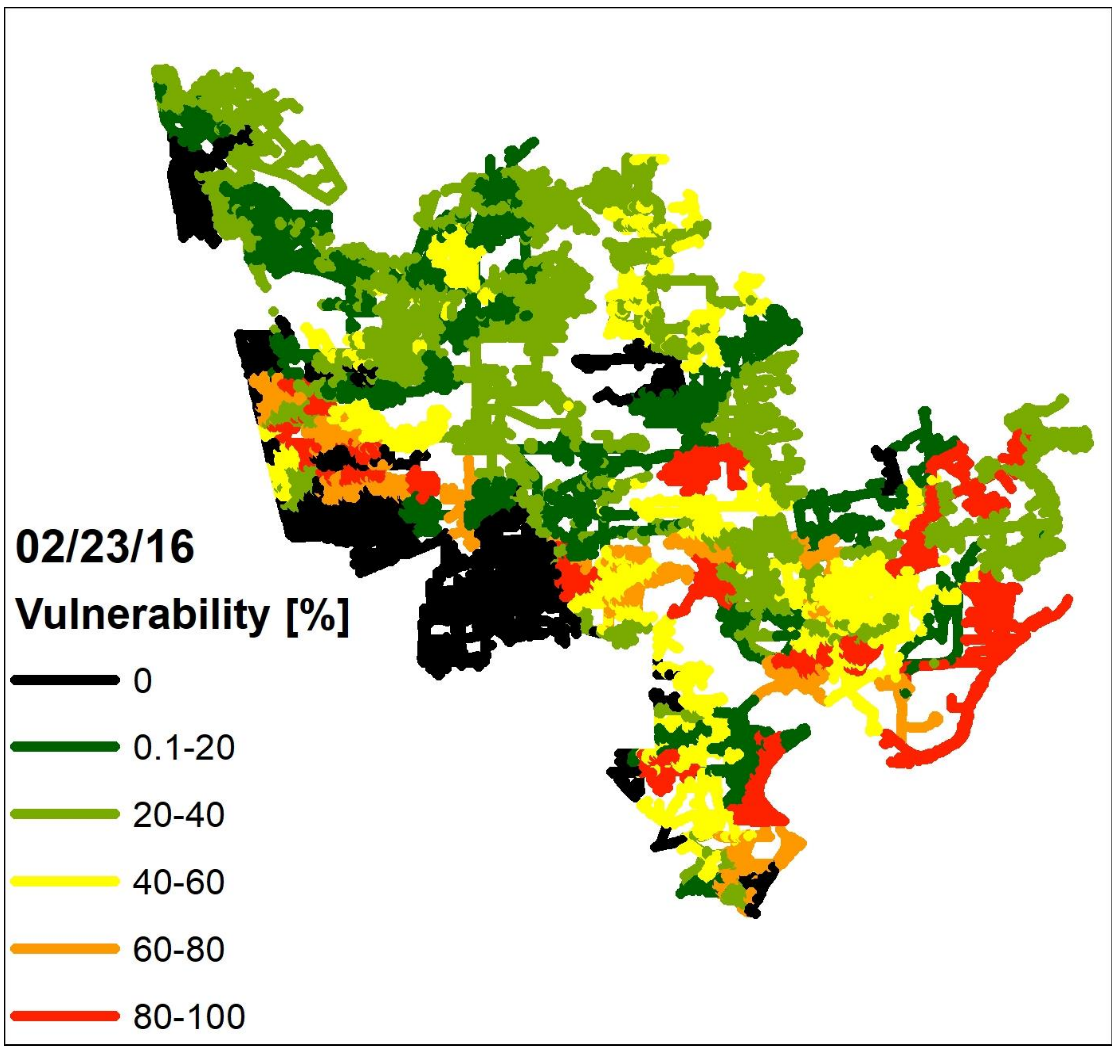

Figure 23 Vulnerability Map for 02/23/2016, reprinted from [53]

risk based only on vegetation indices, performance is limited. Further investigation could be conducted with the larger dataset to test this hypothesis.

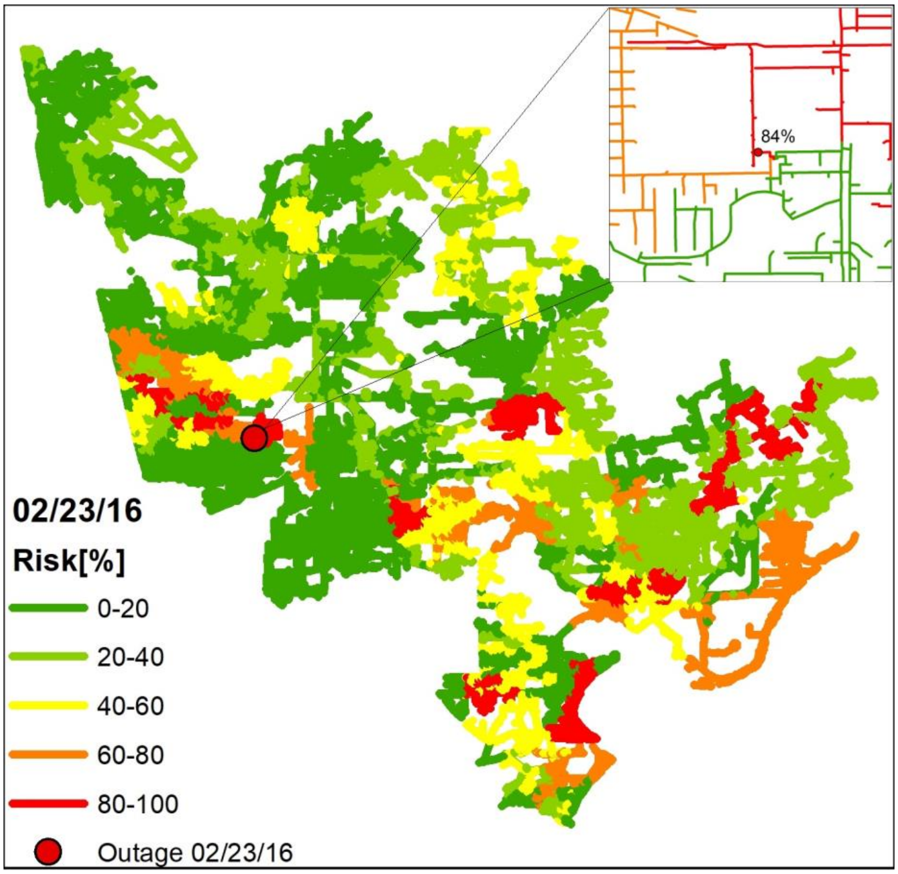

Figure 24 Risk Map for 02/23/2016, reprinted from [53]

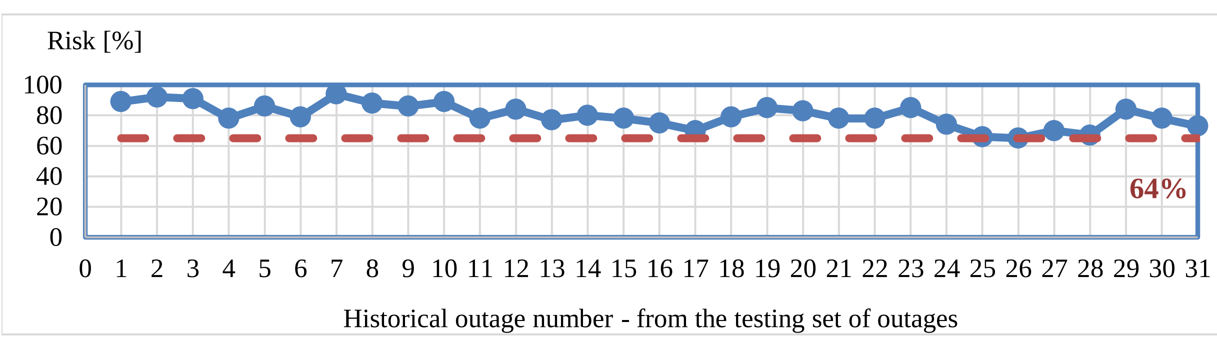

Figure 25 Calculated risk at the end of training for the outages that occurred at the end of 2015 and beginning of 2016, reprinted from [53]

An example of the tree trimming schedule developed for one quarter is presented in Fig. 26. The zones with different colors (not black) represent the areas of the network that need to be trimmed in the selected quarter. These zones change every quarter. The areas that need to be trimmed sooner are represented with red while the areas that need to be trimmed later are represented with green.

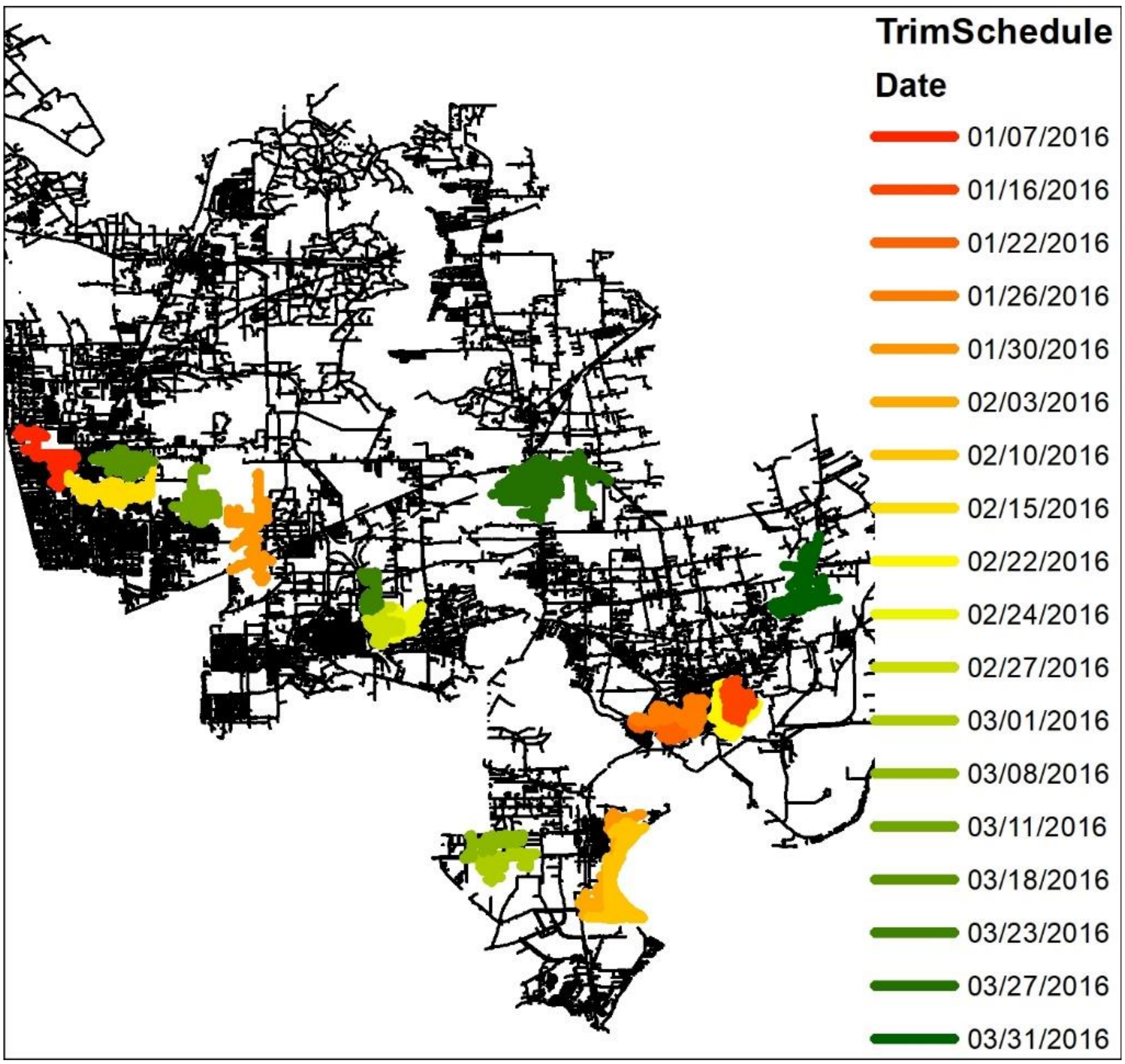

Figure 26 Quarterly Tree Trimming Schedule, reprinted from [53]

The overall outage risk for the selected quarter is calculated as follows:

$$R = \sum_{t=1}^{T} \frac{1}{N} \sum_{n=1}^{N} R_{n,t} \tag{25}$$

The optimal tree trimming schedule reduced the overall outage risk of the network for the period of three months by 32.8%. In addition, the reactive tree trimming total cost described in Sec. 8.4 was predicted to be decreased by 27.2%.

8.6 Conclusion

In this section, we validated the hypothesis as stated at the beginning of the dissertation for the specific application of predictive vegetation risk management for optimal tree trimming scheduling. We also illustrated that this work provides several contributions:

- To improve risk predictions, a variety of data sources are used: historical weather and weather forecast data, various vegetation indices and high-resolution imagery data, and historical utility records about outages and maintenance. The integration and correlation of this data provides a data framework capable of ingesting a large amount of interdisciplinary data, and spatiotemporally correlating it for the extraction of measured variables for all historical events.
- A spatiotemporal model for correlating a variety of data in time and space is developed, which provides real-time generation of predictive risk maps for the assessment of the vegetation around the distribution feeders.
- An analytical approach is introduced for vegetation risk management based on a GCRF, which takes into account both the spatial and the temporal configuration of the network and past events to improve the prediction performance in terms of prediction accuracy and robustness to bad and missing data.

- An optimized, cost-effective dynamic tree trimming scheduler is developed to minimize the overall risk of the network while maintaining the economic investment in periodic tree trimming. The unique benefits of this approach are demonstrated on an actual utility distribution network.

CHAPTER IX

APPLICATION TO INSULATION COORDINATION*

9.1 Introduction

In this chapter, we describe the application to transmission line insulation coordination, based on the predictive three-level framework described in Chapters V-VII. This chapter also serves the purpose of verifying the hypothesis by testing the accuracy of the prediction algorithm, and quantifying the improvements in network reliability achieved based on optimal asset management.

Our implementation provides a new predictive framework for insulator asset maintenance scheduling that combines sensor monitoring data with sets of weather, lightning, vegetation, and GIS data. Instead of statistically estimating failure rates, the data is used to train a prediction model based on linear regression. The state of the network assets is automatically updated, resulting in dynamic risk maps allowing optimized scheduling of assets based on dynamically unfolding risk assessment. The maintenance schedule is created whenever a new set of measurements becomes available and component state is automatically updated over time. The advantages of this framework are illustrated using the intelligent monitoring and maintenance scheduling for transmission line insulators, and the development of the optimal LSA placement strategy.

* This section is in part a reprint with permission of the material in the following papers: (1) M. Kezunovic, T. Dokic, "Predictive Asset Management Under Weather Impacts Using Big Data, Spatiotemporal Data Analytics and Risk Based Decision-Making," 10th Bulk Power Systems Dynamics and Control Symposium – IREP'2017, Espinho, Portugal, August 2017.

9.2 Data

The first step is to identify all the parameters of interest for the development of the predictive risk model. There are six groups of parameters: insulator physical characteristics, insulator deterioration group and in-field measurements, weather, lightning, other environmental factors, and historical network data. Table 10 lists all the parameters in the raw data.

9.2.1 Spatial data analytics

Spatial correlation of data is done using ESRI's ArcGIS platform [101]. The geospatial data model is presented in Fig. 27. Utilities maintain a geodatabase with the locations of all towers, substations, and lines. These are typically stored as shapefiles. Based on the network geodata, the

Table 10 List of parameters, reprinted from [19]

Historical Network Data	**In-field Measurements**	**Weather**	**Lightning**
Outage Reports	Leakage Current Magnitude	Temperature	Peak Current
Maintenance Orders	Flashover Voltage	Humidity	Polarity
Replacement Orders	Electric Field Distribution	Pressure	Type of Lightning
Insulator Physical Characteristics	Corona Discharge Detection	Wind Parameters (speed, direction, gust)	**Other Environmental Parameters**
Surge Impedances of Towers and Ground Wires	Infrared Reflection Thermography	Pollution (sodium chloride)	Vegetation Index (presence and canopy height)
Footing Resistance	Visual Inspection Reports	UV index	Elevation
Component BIL	Radio Interference Voltage	Precipitation	Soil

area of interest for correlated weather data can be selected. This area is then split into 1 km blocks, and all weather parameters are interpolated to the locations of these blocks. The final weather data contains a set of shapefiles with polygons where each time step has one shapefile assigned to it.

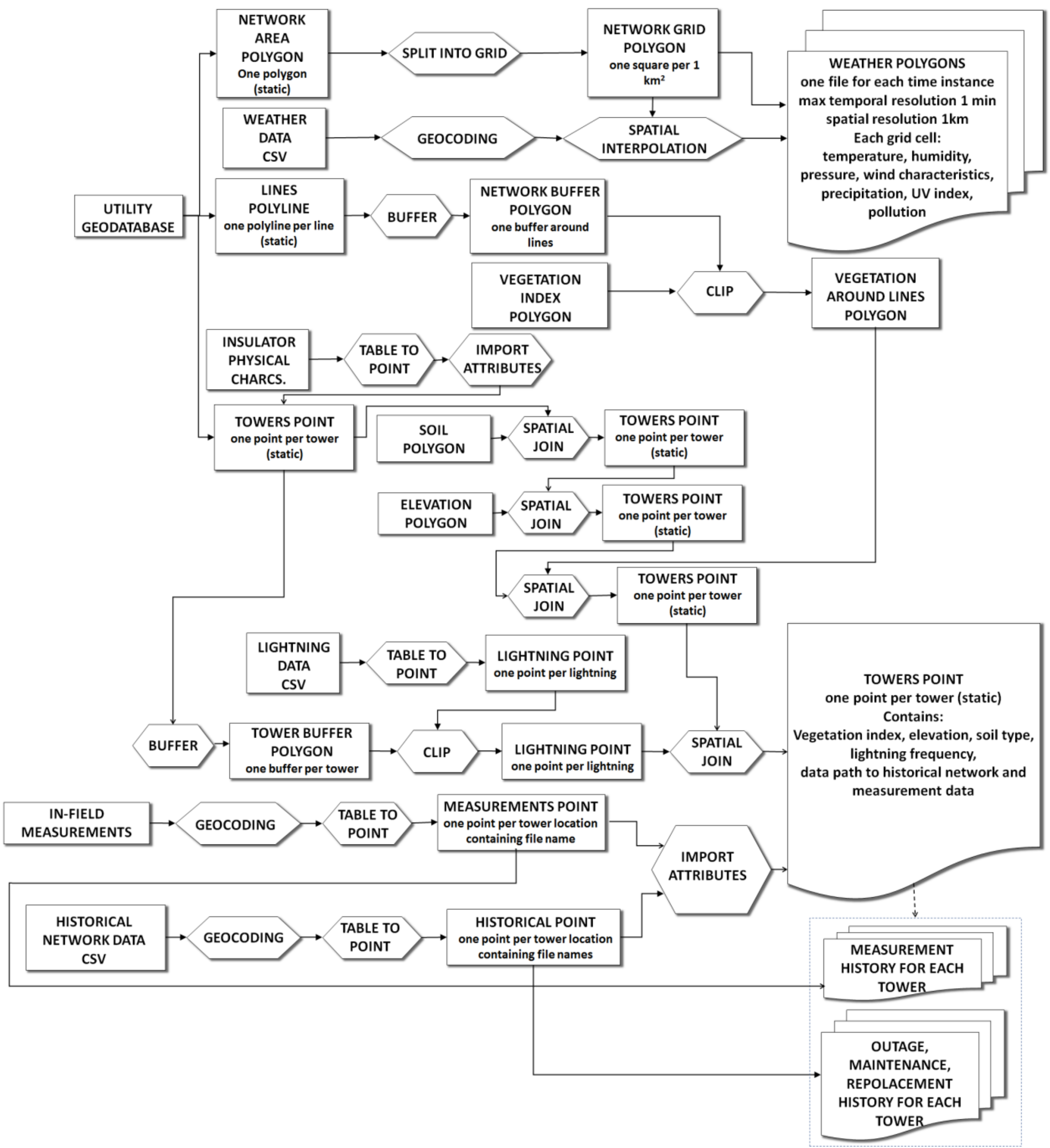

Figure 27 Spatial Correlation of Data, reprinted from [19]

Vegetation data is clipped to a buffer around the lines. This data identifies the parts of a circuit that have tree coverage and are not likely to have lightning-caused outages. The vegetation, elevation, soil, and lightning frequency data is added to the tower shapefile one by one, performing a spatial join to extract the features of the selected file that are closest to the tower point features. Insulator physical characteristics, in-field measurement locations, and historical network data are already geocoded to the tower points, so their attributes can simply be added to the tower attribute table. In-field measurements and historical network data (including historical outage, maintenance, and component replacement data) have temporal components. Thus, for each tower, a pointer to the location of the historical file on the disk is created and added to the attribute table.

The final product of spatial correlation of data are the following datasets:

- Weather Dataset: contains one file for each time step. Every file is a shapefile containing polygons associated with the locational weather parameters.
- Tower Dataset: contains all the parameters projected to the tower location as a point. In addition, the tower dataset contains the link to the historical dataset repository.
- Historical Dataset: contains no geospatial data in its raw form. Each tower in the network has a set of four files inside historical dataset that lists all the events of interest that occurred in the tower’s history. These files are then associated with the geo-location of the towers, which transfers them into a geospatial dataset.

Spatial correlation of data is performed only once as a part of preprocessing. After the initial setup of database, the information is automatically updated with every time step, as described in the following section 9.9.2.

9.9.2 Temporal data analytics

Fig. 28 demonstrates the steps to achieve temporal correlation of data. Different datasets are collected in different time zones. The UTC time standard has been chosen, and all the temporal data is converted to the UTC time zone. The lightning data needs to be temporally correlated with 1) weather data, by interpolating weather parameters at the time instance of the lightning strike, and 2) historical outage data, by identifying which lightning strike corresponds to the historical lightning outage.

The following actions are performed automatically as concurrent processes:

- Lightning Impact: After each lightning strike, changes are applied to the lightning performance characteristics of the insulator, and the insulator state is updated accordingly.

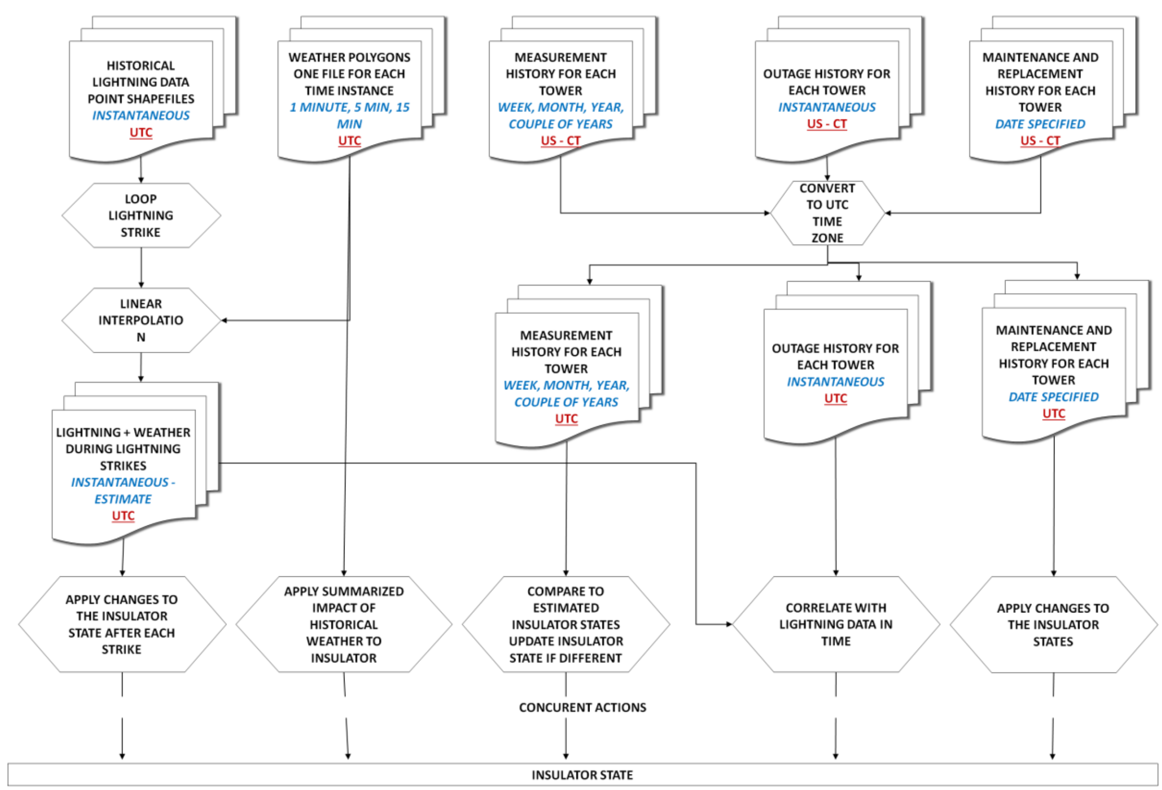

Figure 28 Temporal correlation of data, reprinted from [19]

- Weather Impact: After each month, the weather impacts on the insulator are summarized and the insulator state is updated.
- Measurement-Based State Update: Whenever a measurement in the field is performed, the predicted state of the insulator is compared to the measured characteristics. If there is a difference, the state of the insulator is calibrated to the measured value.
- Outage Impact: Based on the collected outage data, the severity of lightning strike impact on the insulator is determined.
- Maintenance/Replacement: Whenever there is a restorative action in the network, the associated insulator state is refreshed to the repaired value in case of maintenance, or new component value in case of replacement.

9.3 Risk for Insulation Coordination

The risk assessment framework used for this research is defined as in Chapter 6.2, Eq. 7 where T is the lightning peak current, Hazard, $P[T]$, is a probability of a lightning strike with intensity T, $P[C|T]$ is the Vulnerability or probability of an insulation total failure if a lightning strike with intensity T occurred, and the Worth of Loss, $u(C)$, is an estimate of financial losses in case of insulation total failure.

Lightning data indicate the probability of a lightning strike, which impacts the probability of a backflashover. The probability of a backflashover is also impacted by weather conditions (temperature, pressure, humidity and precipitation). If there is a backflashover, the probability of a total component failure (a situation where insulation is significantly damaged and needs to be replaced) is examined. Not every flashover will cause insulation total failure, so the probability of a flashover and probability of insulation total failure are expressed separately and then combined within the overall risk framework. Due to component failure, some losses are expected to be imposed.

9.3.1 Hazard

The probability of a lightning strike is estimated based on historical lightning data in a radius around the affected components. Historical data for a period of 10 years were used. For each node, the lightning frequency is calculated as:

$$LD_i = \frac{L_A}{L_T} \tag{26}$$

where L_A is the number of lightning strikes in the area within a radius of 100 m around the node and L_T is the number of lightning strikes in the total area of the network.

9.3.2 Vulnerability

For the prediction of network vulnerability levels, the GCRF algorithm is used [78]. The advantages of this algorithm are the capability to model the network as interconnected graph with assigned geographical locations and time reference, and the capability to model the interdependencies between different nodes in the network. Input variables include: lightning peak current, lightning polarity, temperature, humidity, pressure, precipitation, temperature variations, UV, and pollution experienced during time step Δt, the presence of a catastrophic event, leakage current magnitude, flashover voltage, corona discharge detection, radio interference voltage, flag for inspection changes detected, *BIL*, and insulator state. The output of the prediction algorithm is the predicted insulator state after the time step Δt. Based on the predicted insulator state, the insulator is placed in one of four groups (as new, weathered, mature, and at risk), and the probability of insulator failure is determined as presented in Fig. 6 in Chapter 4.3.

9.4 Optimal Maintenance Scheduling

The purpose of a maintenance scheduler is to provide a balance between system reliability and maintenance costs. The scheduler is trying to maximize the risk reduction for a system while minimizing the expenses of insulator replacement and maintenance. The available maintenance actions are classified into three groups: do nothing, perform maintenance, or replace a component. For each time instance, every insulator in the network can be assigned one of these three values. To reduce the number of permutations, only insulators that have a risk value higher than 60% are considered for maintenance, and those that have a risk higher than 80% are considered for replacement. The rest of the network is assigned the "do nothing" action. The maintenance and replacement actions are varied in the selected insulator set until the best maintenance plan is found.

The optimal solution for the maintenance plan is determined by solving the following optimization problem that maximizes the system risk reduction:

$$\max\left[\sum_{a=1}^{N} \Delta \mathrm{R}_{\mathrm{M}}(a) SM(a) + \sum_{a=1}^{N} \Delta \mathrm{R}_{\mathrm{R}}(a) SR(a)\right] \tag{27}$$

with the following constraints:

$$\sum_{a=1}^{N} SM(a) MC(a) \le MA$$

$$\sum_{a=1}^{N} SR(a) RC(a) \le RA \tag{28}$$

$$SM(a) + SR(a) \le 1, a = 1,2,\ldots,N$$

where a is a selected insulator, N is a total number of insulators in the network, $\Delta RM(a)$ is a reduction in risk for an insulator that was under maintenance, $\Delta RR(a)$ is a reduction in risk for an insulator that was replaced. $SM(a)$ is 0 if there was no maintenance action and 1 if there was maintenance; $SR(a)$ is 0 if there was no replacement action and 1 if there was replacement, $MC(a)$ is a cost of maintenance for a component a, MA is a total allocated maintenance fund, $RC(a)$ is the cost of replacement of component a, RA is a total allocated replacement fund.

9.5 Optimal Location of LSA

The global state of risk function is constructed as an arithmetic mean of the individual state of risk for each network component, and summarized over time:

$$R = \frac{1}{N}\sum_{n=1}^{N} R_n \tag{29}$$

where R is a total risk for the entire network, N is the total number of towers in the network, and R_n is the individual risk for tower n. The optimization algorithm maximizes the global state of risk reduction by setting LSA positions as independent variables:

$$\max\left\{R = \frac{1}{N}\sum_{n=1}^{N} \Delta R_n \cdot F_n\right\}$$
$$F_n = \begin{cases} 0, & \text{no LSA} \\ 1, & \text{LSA installed} \end{cases} \tag{30}$$

where ΔR_n is a risk reduction on a tower n after installation of LSA. The available budget for the LSA installation is considered to be limited, adding an economic constraint:

$$\sum_{n=1}^{N} F_n \cdot C_n \le TC \tag{31}$$

where C_n is a cost of installation of LSA on tower n, and TC is a total budget dedicated to the LSA installations.

9.6 Evaluation and Results

The method has been simulated and tested on a 36 substation, 65 transmission lines section of a network, with a total of 1590 towers. The data coming from three ASOS weather stations [30] located in the vicinity of the network was used, and lightning data was obtained from the NLDN [31]. Weather forecast data used in this study was downloaded from the NDFD [32].

In Fig. 29, an example of a Hazard map is presented. The Hazard map is created based on the interpolated weather data and presented as a polygon grid, where each block has an assigned

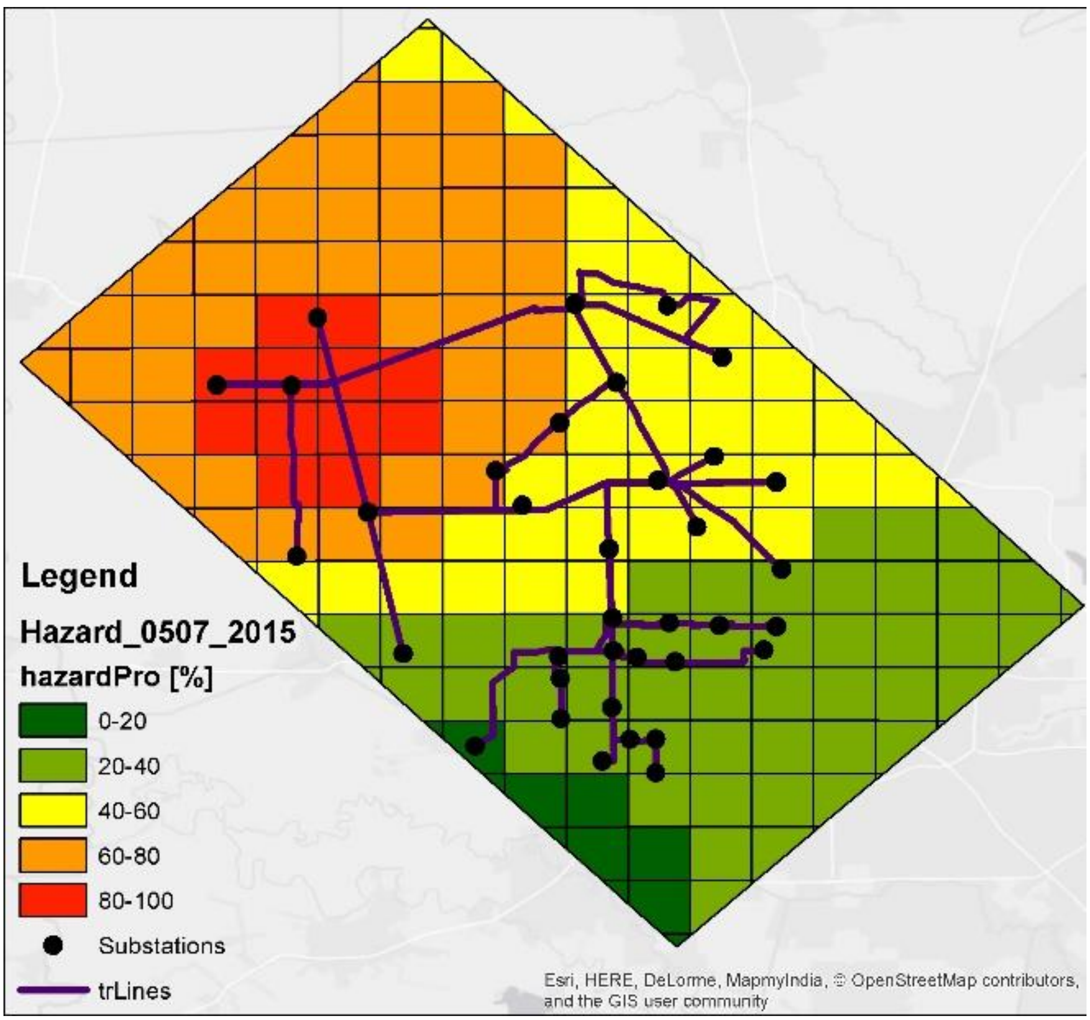

Figure 29 Weather Hazard Map, reprinted from [139]

hazard probability. An example of a Vulnerability map is presented in Fig. 30. Each tower has a vulnerability value assigned to it. The Vulnerability value represents the probability of an insulator failure if the presented Hazard in Fig. 29 has occurred.

In Fig. 31, an example of a risk map for one time instance is presented. The advantage of our method is that risk maps are generated continuously over time. At each moment new data is available; the appropriate risk map is assigned based on the current weather forecast and current conditions of network assets.

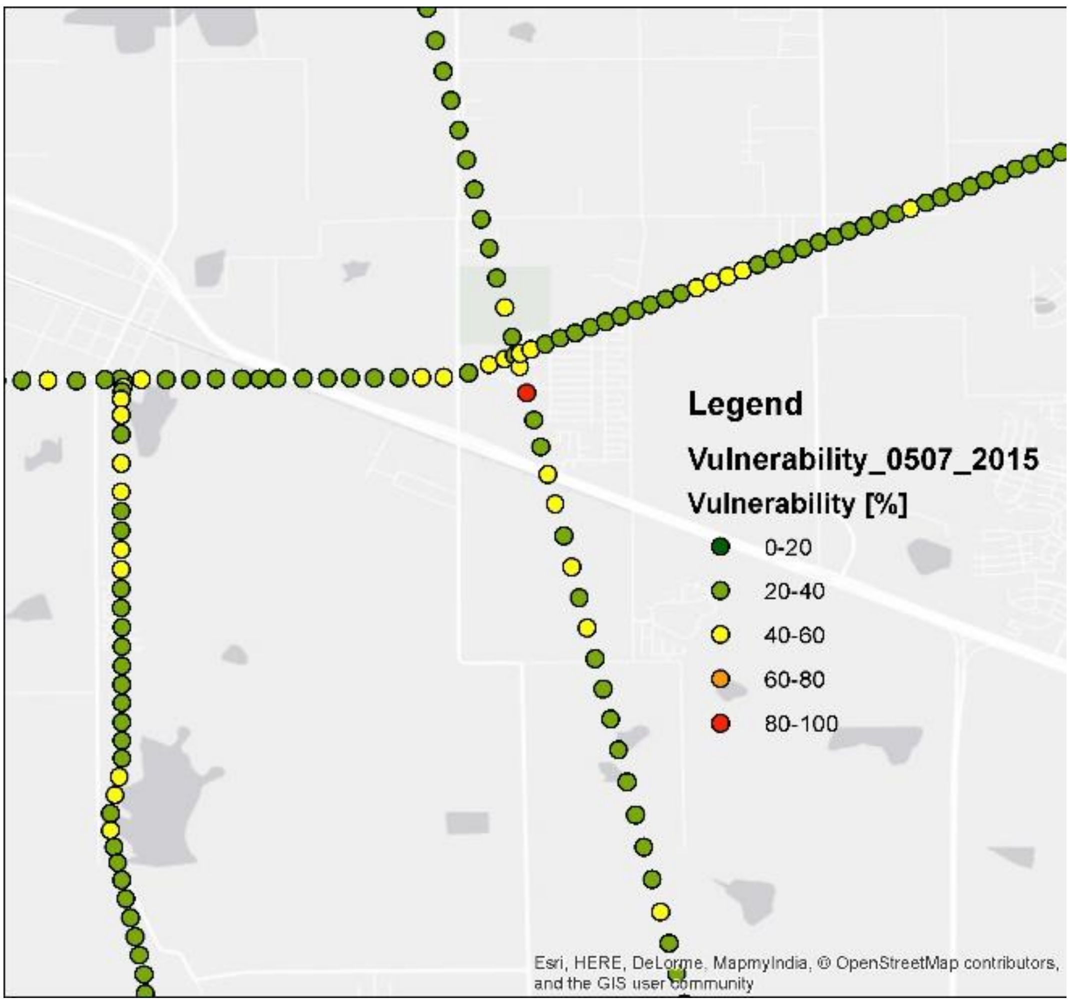

Figure 30 Tower Vulnerability Map, reprinted from [139]

Based on the overall risk map created for a period of one year and associated economic cost, the optimal maintenance plan is presented in Table 11. The presented maintenance plan is expected to reduce overall risk by 56% during one year of application. With this method, the asset maintenance schedule is determined dynamically and it differs based on different environmental impacts on the network. The dynamic scheduler is constantly learning and adjusting the maintenance schedule to include the impact of all the events in the network.

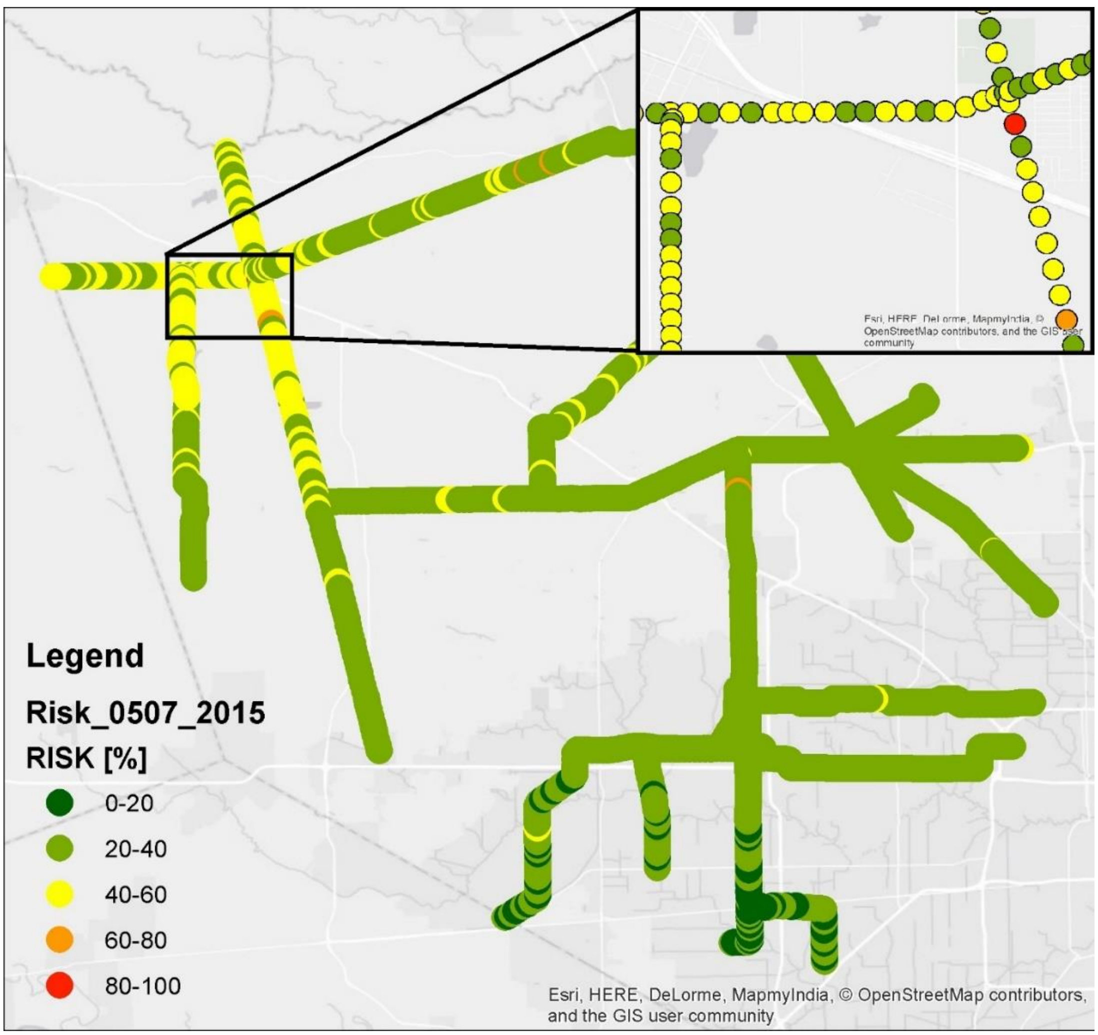

Figure 31 Risk Map of the Network, reprinted from [139]

The recommended number of line surge arresters (LSAs) is calculated to be 264, and the optimal locations of the LSAs in terms of risk reduction are presented in Fig. 32. The presented configuration of LSAs is expected to reduce the overall risk by 72%. This kind of result could help utilities make decisions about the installation of LSAs in an economically efficient way.

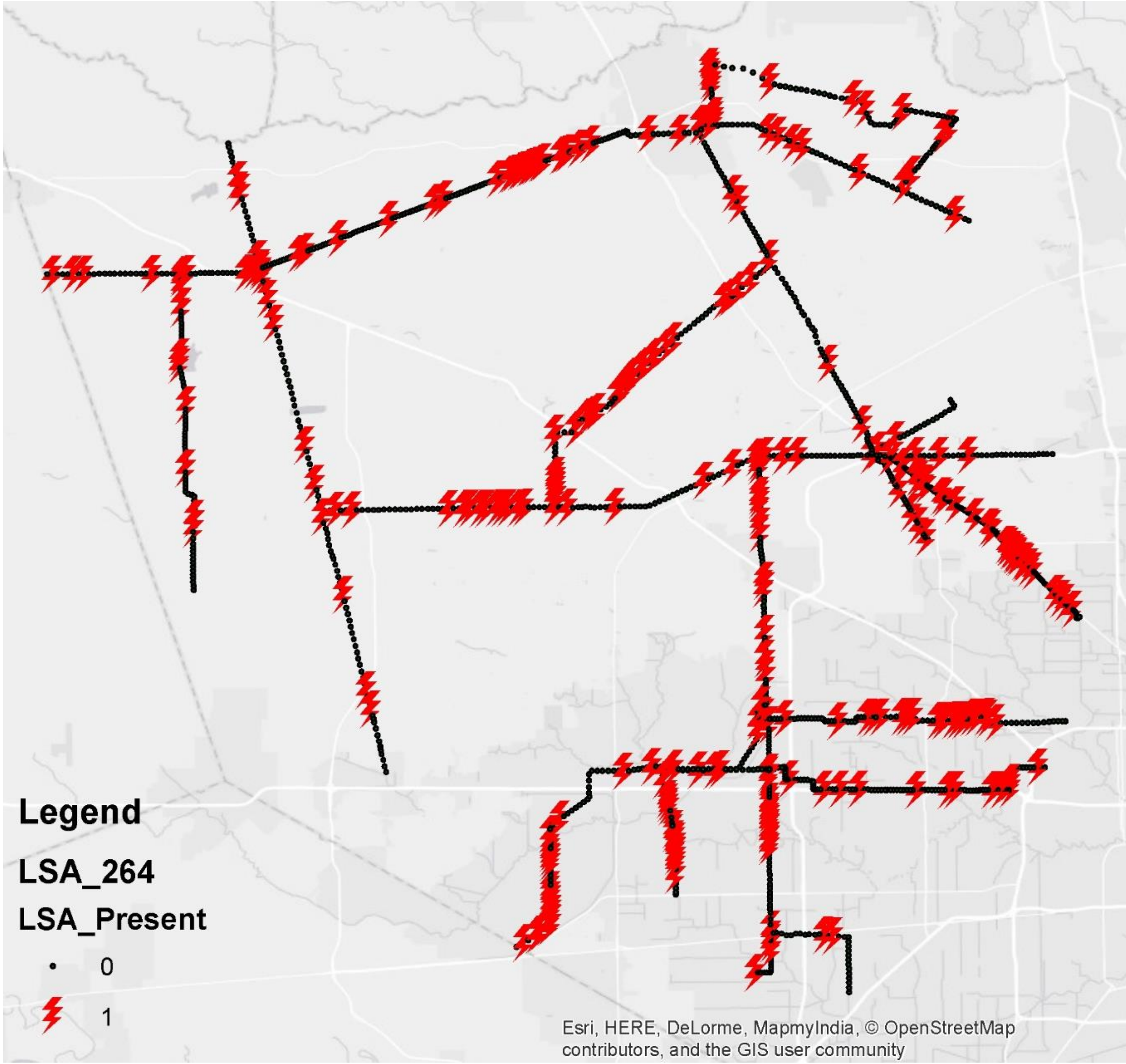

Figure 32 Locations of 264 Line Surge Arresters, reprinted from [139]

Table 11 Optimal maintenance schedules, reprinted from [19]

Time step (month)	Insulator ID	Type of action	System Risk Reduction [%]
1	1528, 924, 949, 321, 1111	M	18.02
	152, 954	R	
2	333, 851, 29, 1374, 854, 376	M	15.13
	34, 641	R	
3	525, 241, 384, 964, 464, 56	M	17.52
	944	R	
4	309, 1191, 1352	M	5.27
	861	R	
5	1506, 1208, 592, 559, 243	M	28.24
	185	R	
6	1389, 1443, 1064, 1009, 345, 127	M	13.13
	528, 74	R	
7	511, 130, 1008	M	10.54
	1181	R	
8	574, 254, 367	M	22.84
	497, 98	R	
9	1435, 1471, 502, 1535, 131	M	26.51
	771, 1313	R	
10	612, 1244, 787	M	7.89
	654	R	
11	217, 70, 369, 137	M	30.02
	184	R	
12	1524, 1475, 1232	M	12.44
	1485, 1501	R	

9.7 Conclusion

This application entails dynamic maintenance scheduling for predictive insulator asset management of geo-spatially and temporally referenced data. The following are the benefits of this application:

- A novel predictive asset management framework is proposed that optimizes the maintenance schedule based on the dynamically created State of Risk.
- The spatial and temporal integration of input data results in locational assessment of asset deterioration over time.
- The model is capable of predicting the future State of Risk based on the GCRF model, which is scalable to include a large number of asset components.
- The prediction model takes into account the spatial interdependencies of the input data, providing the additional knowledge for more precise prediction of the State of Risk.
- The method combines sensor data used for condition monitoring with additional data obtained from sources outside of the electric utility, such as weather data, which gives a more precise assessment of the asset aging.
- The optimal LSA placement strategy based on prediction of network state of risk for lightning related outages has been developed.

CHAPTER X

CONCLUSIONS

10.1 Introduction

Weather impacts are the main cause of outages in power systems, causing utilities and customers to lose millions of dollars each year, and sometimes even putting the public in danger. The advancements in Big Data and data analysis have opened a new path for the development of predictive applications in power systems. The use of Big Data for the evaluation of weather impacts and the mitigation of consequences is an emerging area of research.

In demonstrating the practical value of the solution, we placed a focus on applications of weather impact on asset and outage management in both transmission and distribution. The goal was to select and apply the most suitable prediction algorithm that can take advantage of the underlying data structure, as well as mitigate problems such as bad and missing data to produce prediction results not feasible before. The hypothesis of this dissertation was that such a solution will be able to provide a high prediction accuracy that will move predictive outage analysis from predicting the number of outages per area or risk for an outage in an area, to the level of prediction where precise risks of outages for components are predicted. This in turn will have a major impact on resiliency, cost reduction and customer satisfaction.

With a judicious selection of prediction algorithm characteristics, we achieved the capability to predict outages on multiple temporal and spatial scales. This contribution has paved the way for accurate localized real-time prediction of outages and asset failures, which is not available to different maintenance strategies in use today. This kind of approach can significantly improve the efficiency and economic impact of asset management because it allows optimized mitigation strategies to be developed ahead of the time, reducing the response time and cost of the resources.

10.2 Contributions

The follow-on achievement of this dissertation is an introduction of a new framework for optimal asset management based on predictive risk assessment that is illustrated with examples for distribution and transmission assets. The following are the specific innovative contributions and related benefits:

- The unified data framework that enables the collection and spatiotemporal correlation of a variety of data sets serving different applications was developed. The study uses a variety of data sources, some collected by the utility such as outage and assets data; and an extensive set of environmental data, such as weather station data, weather forecast, vegetation, and lightning. The study developed data-driven techniques using all the measured parameters that surround the event of interest.
- The temporal and spatial interdependencies between components and events in the network are leveraged for improving the prediction algorithm's accuracy and ability to deal with bad and missing data. The prediction is done on multiple spatial and temporal scales. This enables the utilization of results for different purposes, ranging from operation to asset management and planning, and can significantly increase the overall value of the data for the utility. This is the first time that the prediction of weather impacts was made on the component level, without averaging the impacts over the area or component type.
- The Gaussian Conditional Random Fields (GCRF) algorithm was used for the prediction of the probability of future outages in the network for the given weather

forecast data. The algorithm shows a high accuracy of prediction by predicting risk of 64% or higher for all the cases of outages in distribution, and over 74% of accuracy for cases in transmission. Three main reasons this algorithm is capable of achieving high performances for these applications are as follow: 1) it is a graph based structured regression capable of modeling all the aspects of the network; 2) Conditional Random Fields enables the extraction of knowledge from the spatial interdependencies of the outputs that improves the accuracy of prediction since the events of interest have higher correlation if they are near each other; 3) fast computation, capable of handling large electric networks with multiple hundreds of thousands of nodes, is achieved by making the feature functions Gaussian. We introduced a study that achieves the integration of the presented GCRF algorithm with multiple different applications in power systems.

- The prediction results are presented on a geographical map in the form of Risk Maps. These maps are created dynamically based on the available weather forecast. The maps provide a precise prediction of risk for every component in the network at the moment in time when the weather forecast was made. Unlike the existing solutions that map the risk in various larger areas over the network, the risk maps developed in this research support multiple different spatial resolutions, capable of mapping the risk for the individual components, compared to existing solutions that average the impacts over a number of components.
- A dynamic asset management system based on optimization was built to reduce the predicted risk of outages and component failure while maintaining predetermined economic investment in periodic asset maintenance. The novelty of this

optimization system is that it uses the outputs of the predictive risk analysis to optimize the decision-making process based on the predicted levels of risk for each individual component in the network.

- The optimization was demonstrated on three different cases: 1) optimal tree trimming scheduling for distribution systems, 2) optimal maintenance/replacement of transmission line insulators, and 3) optimal placement of Line Surge Arresters. The method is applied to real utility data and the prediction performance in a real-life setting is evaluated.

10.3 Future work

Due to continuous advancements in instrumentation and measurements, more and more new data sets are being collected every day. The study in this dissertation leveraged a large amount of data that was available at the time the research was conducted. There are constantly new datasets becoming available, and it is expected that this trend will continue in the future.

For example, more and more 3-dimensional geographical data is collected by different organizations thanks to LIDAR technology and various drone surveys. New algorithms will have to be developed to extract all the necessary information from the 3-dimensional data and incorporate it into prediction algorithms. Methods such as computer vision could be used for this purpose.

Advancements in weather forecasting are creating better prediction on multiple temporal scales with every new discovery in the field. One aspect that was not analyzed in our study, and that would be of potential use for improvement of the prediction algorithm's capabilities, is how uncertainty in the weather forecast affects the outage prediction. This could result in the creation of a family of prediction algorithms that are capable of providing the best possible accuracy for each time horizon of weather forecast. In addition to uncertainties in the weather forecast, there are other uncertainties introduced in the study during the preprocessing and spatiotemporal correlation of data, such as error due to spatial interpolation of weather measurements, error due to temporal misplacement of vegetation measurements, etc. It would be of great benefit to develop a comprehensive uncertainty study that can model different errors and their impact on the prediction result, which will correlate the range of possible output values based on these uncertainties.

With the addition of new developing data sets it will be interesting to see how good the prediction performances of algorithms become in the future. The research in this dissertation is the first step in creating precise component-based prediction of weather impacts on electric transmission and distribution. However, the accuracy of the prediction is highly dependent on the availability, quality and spatiotemporal resolution of data sets. It is also dependent on the specific configuration of prediction algorithms that could always be improved to match the problem's specific data structure. In this work a single data structure was used, even though different data sets come from different types of networks. This could be extended to a multilayer solution that combines different data structures. Each data structure would have to be the best match for the specific data set of interest. The future should focus on demonstrating how far the capabilities of predictive algorithms can go when estimating weather impacts on power systems.

REFERENCES

[1] Eaton, “Blackout and Power Outage Tracker,” 2019 [Online] Available: https://switchon.eaton.com/blackout-tracker

[2] We Energies, “Power outage causes and restoration,” 2019 [Online] Available: https://www.we-energies.com/outages_safety/reporting/outages.htm

[3] M. Panteli, P. Mancerella, “Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies,” Electric Power Systems Research, Vol. 127, pp. 259–270, 2015.

[4] Q. Yan, T. Dokic, M. Kezunovic, “Predicting Impact of Weather Caused Blackouts on Electricity Customers Based on Risk Assessment,” IEEE Power and Energy Society General Meeting, Boston, MA, July 2016.

[5] T. Dokic, M. Kezunovic, "Real-Time Weather Hazard Assessment for Power System Emergency Risk Management," CIGRE US National Committee 2017 Grid of the Future Symposium, Cleveland, OH, October 2017.

[6] E. Britt, “Hurricanes Harvey and Irma: Electric Industry Impacts, Restoration, and Cost Recovery,” Infrastructure, Vol. 57, No. 1, Fall 2017.

[7] M. Panteli, P. Mancarella, “Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies,” Electric Power Systems Research, Vol. 127, October 2015, pp. 259-270.

[8] P. Hines, et al., “Trends in the History of Large Blackouts in the United States,“ IEEE Power and Energy Society General Meeting, IEEE, 2008.

[9] A. Kenward, and U. Raja, “Blackout: Extreme Weather, Climate Change and Power Outages,”2014 [Online] Available: http://assets.climatecentral.org/pdfs/PowerOutages.pdf

[10] M. Kezunovic, Z. Obradovic, T. Dokic, S. Roychoudhury, "Systematic Framework for Integration of Weather Data into Prediction Models for the Electric Grid Outage and Asset Management Applications," The Hawaii International Conference on System Sciences - HICSS, Waikoloa Village, Hawaii, January 2018.

[11] A. Baggini, “Handbook of Power Quality,” Wiley, July 2008. ISBN: 978-0-470-75423-8

[12] North American Electric Reliability Council, “Technical Analysis of the August 14, 2003, Blackout: What Happened, Why, and What Did We Learn?” Report to the NERC Board of Trustees by the NERC Steering Group, July 2004.

[13] U.S.-Canada Power System Outage Task Force, “Final Report on the Implementation of the Task Force Recommendations” September 2006.

[14] NEI Electric Power Engineering, New Hampshire December 2008 Ice Storm Assessment Report, Oct. 2009.

[15] R. J. Campbell, “Weather-Related Power Outages and Electric System Resiliency,” Congressional Research Service, Aug. 2012.

[16] Federal Energy Regulatory Commission Untied States Government, “Utility Vegetation Management Final Report,” Mar. 2004.

[17] T. Dokic, P.-C. Chen, M. Kezunovic, "Risk Analysis for Assessment of Vegetation Impact on Outages in Electric Power Systems," CIGRE US National Committee 2016 Grid of the Future Symposium, Philadelphia, PA, October-November 2016.

[18] Carroll Electric Cooperative, “Cost Analysis for Integrated Vegetation Management Plan,” 20 April 2010. [Online] Available: http://www.carrollecc.com/files/pdf/cecc_finley_cost_study.pdf

[19] M. Kezunovic, T. Dokic, "Predictive Asset Management Under Weather Impacts Using Big Data, Spatiotemporal Data Analytics and Risk Based Decision-Making," 10th Bulk Power Systems Dynamics and Control Symposium – IREP'2017, Espinho, Portugal, August 2017.

[20] T. Dokic, P. Dehghanian, P.-C. Chen, M. Kezunovic, Z. Medina-Cetina, J. Stojanovic, Z. Obradovic "Risk Assessment of a Transmission Line Insulation Breakdown due to Lightning and Severe Weather," HICCS – Hawaii International Conference on System Science, Kauai, Hawaii, January 2016.

[21] R. S. Gorur, et al., "Utilities Share Their Insulator Field Experience," T&D World Magazine, Apr. 2005, [Online] Available: http://tdworld.com/overhead-transmission/utilities-share-their-insulator-field-experience

[22] P. R. Ilich, R. Riddles, W. Haak, R. Frowd, "Application of CIM model for enterprise wide power system model for Planning, Protection and Operations, " Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, Pittsburgh, PA, July 2008.

[23] A. Abiri-Jahromi, M. Parvania, F. Bouffard, M. Fotuhi-Firuzabad, "A Two-Stage Framework for Power Transformer Asset Maintenance Management—Part I: Models and Formulations, " Power Systems, IEEE Transactions on, Vol. 28, No. 2, pp. 1395-1403, October 2012.

[24] T. Jinrui, Y. Xianggen, Z. Zhe, "Modeling Technology in Traveling- Wave Fault Location, " TELEKOMNIKA, Vol. 11, No. 6, pp. 3333-3340, Jun. 2013.

[25] J. R. Marti "Accurate modeling of frequency-dependent transmission lines in electromagnetic transient simulations, " IEEE Trans. Power Apparatus and Syst., Vol. PAS-101, No. 1, pp. 147-157, Jan. 1982.

[26] M. Rui, H. Renmu, W. Peng, "A Multiobjective Operation Planning Model With Transmission Constraints and Load Characteristic of Distribution System, " Transmission and Distribution Conference and Exhibition, 2005/2006 IEEE PES, Dallas, TX, May 2006.

[27] S. Schlegel, R. Schwerdfeger, D. Westermann, A. Kuester, "Modelling of transmission system operation in Europe for enhanced security of supply, " EUROCON, 2013 IEEE, Zagreb, July 2013.

[28] National Oceanic and Atmospheric Administration, "Radar Data in the NOAA Big Data Project," [Online] Available: https://www.ncdc.noaa.gov/dataaccess/radar-data/noaa-big-data-project

[29] National Oceanic and Atmospheric Administration, "Satellite Data," 2017 [Online] Available: https://www.ncdc.noaa.gov/data-access/satellite-data

[30] National Oceanic and Atmospheric Administration, "Automated Surface Observing System (ASOS)," [Online] Available: https://www.ncdc.noaa.gov/dataaccess/land-based-station-data/land-baseddatasets/automated-surface-observing-system-asos

[31] Vaisala, "National Lightning Detection Network – Technical Specification," 2017 [Online] Available: http://www.vaisala.com/en/products/thunderstormandlightningdetectionsystems/Pages/NLDN.aspx

[32] National Digital Forecast Database (NDFD) Tkdegrib and GRIB2 DataDownload and ImgGen Tool Tutorial, NWS, NOAA. 2017 [Online] Available: http://www.nws.noaa.gov/ndfd/gis/ndfd_tutorial.pdf

[33] Elliott, Lee F., David D. Diamond, C. Diane True, Clayton F. Blodgett, Dyan Pursell, Duane German, and Amie Treuer-Kuehn. 2014. Ecological Mapping Systems of Texas: Summary Report. Texas Parks & Wildlife Department, Austin, Texas.

[34] TNRIS, “Maps & Data,” [Online] Available: www.tnris.org/get-data

[35] TNRIS, Lidar Data, [Online] Available: https://tnris.org/news/2017-06-12/tnris-lidar-data-now-available-download/

[36] J. Endrenyi, et al., “The present status of maintenance strategies and the impact of maintenance on reliability,” IEEE Trans. on Power Sys, Nov. 2001, pp. 638-646.

[37] Nelson, W. “Analysis of accelerated life test data-Part I: The Arrhenius model and graphical methods,” IEEE Transactions on Electrical Insulation, Vol. 4, pp. 165-181, 1971.

[38] J. Endrenyi et al., “The present status of maintenance strategies and the impact of maintenance on reliability,” IEEE Transactions on Power Systems, Vol. 16, No. 4, pp. 638-646, Nov 2001.

[39] S. Natti, M. Kezunovic, "Assessing Circuit Breaker Performance Using Condition-based Data and Bayesian approach," Electric Power Systems Research, Vol. 81, No. 9, pp. 1796-1804, September 2011.

[40] Yeddanapudi, S. R. K., et al. “Risk-based allocation of distribution system maintenance resources,” IEEE Transactions on Power Systems, Vol. 23, No. 2, pp. 287-295, 2008.

[41] Abiri-Jahromi, et al., "An efficient mixed-integer linear formulation for long-term overhead lines maintenance scheduling in power distribution systems." IEEE Transactions on Power Delivery, Vol. 24, No. 4, pp. 2043-2053, 2009.

[42] Liu, Nian, et al., "Asset Analysis of Risk Assessment for IEC 61850-Based Power Control Systems—Part I: Methodology." IEEE Transactions on Power Delivery Vol. 26, No. 2, pp. 869-875, 2011.

[43] Liu, Nian, et al., "Asset analysis of risk assessment for IEC 61850-based power control systems—Part II: Application in substation." IEEE Transactions on Power Delivery Vol. 26, No. 2, pp. 876-881. 2011.

[44] Qiu, Jian, et al. "Nonparametric regression-based failure rate model for electric power equipment using lifecycle data." IEEE Transactions on Smart Grid, Vol. 6, No. 2, pp. 955-964, 2015.

[45] Liu, Xindong, et al. "Risk assessment in extreme events considering the reliability of protection systems." IEEE Transactions on Smart Grid Vol. 6, No. 2, pp. 1073-1081, 2015.

[46] P. Wang, and Roy Billinton. "Reliability cost/worth assessment of distribution systems incorporating time-varying weather conditions and restoration resources." IEEE Transactions on Power Delivery, Vol. 17, No. 1, pp. 260-265, 2002.

[47] L. Treinish et al., “Operational utilization and evaluation of a coupled weather and outage prediction service for electric utility operations,” in Proc. 2nd Conf. Weather Climate New Energy Economy, Seattle, WA, USA, Jan. 2011.

[48] D. Lubkeman, and D. E. Julian. "Large scale storm outage management." Power Engineering Society General Meeting, IEEE, Denver, CO, 2004.

[49] G. Li, , et al. "Risk analysis for distribution systems in the northeast US under wind storms." IEEE Transactions on Power Systems, Vol. 29, No. 2, pp. 889-898, 2014.

[50] D. Yates, et al. "Stormy weather: Assessing climate change hazards to electric power infrastructure: A Sandy case study." IEEE Power and Energy Magazine, Vol. 12, No. 5, pp. 66-75, 2014.

[51] M. Panteli, et al. "Power System Resilience to Extreme Weather: Fragility Modelling, Probabilistic Impact Assessment, and Adaptation Measures." IEEE Transactions on Power Systems, Vol. 32, No. 5, pp. 3747 – 3757, 2016.

[52] C. Chen, et al., "Modernizing Distribution System Restoration to Achieve Grid Resiliency Against Extreme Weather Events: An Integrated Solution," Proceedings of the IEEE, Vol. 105, No. 7, pp. 1267 – 1288, 2017.

[53] T. Dokic, M. Kezunovic, "Predictive Risk Management for Dynamic Tree Trimming Scheduling for Distribution Networks," IEEE Transactions on Smart Grid, Early Access, September 2018.

[54] D. T. Radmer, P. A. Kuntz, R. D. Christie, S. S. Venkata, R. H. Fletcher, "Predicting Vegetation-Related Failure rates for Overhead Distribution Feeders," IEEE Transactions on Power Delivery, Vol. 17, No. 4, pp. 1170-1174, October 2002.

[55] R. Sievanen and T. E. Burk, "Adjusting a process-base growth model for varying site conditions through parameter estimation," Can. J. Forest Res., Vol. 23, No. 9, pp. 1837–1851, Sept. 1993.

[56] D. C. Hamlin and R. A. Leary, "An Integral—Differential equation model of tree height growth," in Proc. IUFRO Conf., Vol. 2, Minneapolis, MN, pp. 683–690, Aug. 1987,

[57] P. A. Kuntz, R. D. Christie, S. S. Venkata, "Optimal Vegetation Maintenance Scheduling of Overhead Electric Power Distribution Systems," IEEE Transactions on Power Delivery, Vol. 17, No. 4, pp. 1164-1169, October 2002.

[58] J. Wingfield, "New York Power Authority Develops Vegetation Management Solution for High-Voltage Transmission Lines with GIS," ArcNews, Vol. 27, No. 1, pp. 35-, ESRI Spring 2005.

[59] P. A. Kuntz, R. D. Christie, S. S. Venkata, "A Reliability Centered Optimal Visual Inspection Model for Distribution Feeders," IEEE Transactions on Power Delivery, Vol. 16, No. 4, pp. 718-723, October 2001.

[60] S. Guikema, R. A. Davidson, H. Liu, "Statistical Models of the Effects of Tree Trimming on Power System Outages," IEEE Transactions on Power Delivery, Vol. 21, No. 3, pp. 1549-1557, July 2006.

[61] Y. Kobayashi, G. G. Karady, G. T. Heydt, R. G. Olsen, "The Utilization of Satellite Images to Identify Trees Endangering Transmission Lines," IEEE Transactions on Power Delivery, Vol. 24, No. 3, pp. 1703-1709, July 2009.

[62] Y. Kobayashi, G. G. Karady, G. T. Heydt, M. Moeller, R. G. Olsen, "Satellite Imagery for the Identification of Interference with Overhead Power Lines," PSerc Report 08-02, January 2008.

[63] B. D. Russell, C. L. Benner, J. Wischkaemper, W. Jewell, J. McCalley, "Reliability Based Vegetation Management Through Intelligent System monitoring," PSerc Report 07-31, September 2007.

[64] J. A. Wischkaemper, C. L. Benner, B. D. Russell, "Electrical Characterization of Vegetation Contacts with Distribution Conductors – Investigation of Progressive Fault

Behavior," Transmission and Distribution Conference and Exposition, Chicago, IL, April 2008.

[65] P.-C. Chen, T. Dokic, N. Stokes, D. W. Goldberg, M. Kezunovic, "Predicting Weather-Associated Impacts in Outage Management Utilizing the GIS Framework," IEEE/PES Innovative Smart Grid Technologies Latin America (ISGT-LA), Montevideo, Uruguay, October 2015.

[66] I. M. Rawi, et al., "Lightning study and experience on the first 500kV transmission line arrester in Malaysia," International Conference on Lightning Protection (ICLP), Shanghai, China, 2014.

[67] W. Sones, S. M. Wong, "Overview on Transient Overvoltages and Insulation Design For a High Voltage Transmission System," International Conference on High Voltage Engineering and Application (ICHVE), New Orleans, LA, 2010.

[68] Z. G. Datsios, et al., "Estimation of the minimum shielding failure current causing flashover in overhead lines of the hellenic transmission system through ATP-EMTP simulations." International Conference on Lightning Protection (ICLP), Vienna, Austria, 2012.

[69] S. T. Mobarakei, T. Sami, B. Porkar, "Back flashover phenomenon analysis in power transmission substation for insulation coordination," 11th Int. Conf. on Environment and Electrical Engineering (EEEIC), Venice, May 2012.

[70] S. Bedoui, et al., "Analysis of lightning protection with transmission line arrester using ATP/EMTP: Case of an HV 220kV double circuit line." 45th International IEEE Universities Power Engineering Conference (UPEC), Cardiff, Wales, UK, 2010.

[71] M. Kezunovic, et al., "Improved Transmission Line Fault Location Using Automated Correlation of Big Data from Lightning Strikes and Fault-induced Traveling Waves," 48th Hawaii Int. Conf. Syst. Sciences (HICSS), Jan. 2015.

[72] T. Sadovic, et al., "Expert System for Transmission Line Lightning Performance Determination", CIGRE Int. Colloq. on Power Quality and Lightning, Sarajevo, Jun. 2012.

[73] Zhang, J., et al. "Application of hourly meteorological records to atmospheric correction factors in insulation coordination under switching impulse voltage." High Voltage Engineering and Application, Int. Conf. on. IEEE, 2008.

[74] R. Shariatinasab, et al., "Probabilistic evaluation of optimal location of surge arresters on EHV and UHV networks due to switching and lightning surges." IEEE Transactions on Power Delivery, Vol. 24, No. 4, pp. 1903-1911, 2009.

[75] J. Stojanovic, et al,. "Semi-supervised learning for structured regression on partially observed attributed graphs," SIAM International Conference on Data Mining, Vancouver, Canada, April 30 - May 02, 2015.

[76] P. R. Ilich, R. Riddles, W. Haak, R. Frowd, "Application of CIM model for enterprise wide power system model for Planning, Protection and Operations, " Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, Pittsburgh, PA, July 2008.

[77] A. Abiri-Jahromi, M. Parvania, F. Bouffard, M. Fotuhi-Firuzabad, "A Two-Stage Framework for Power Transformer Asset Maintenance Management—Part I: Models and Formulations, " Power Systems, IEEE Transactions on, Vol. 28, No. 2, pp. 1395-1403, October 2012.

[78] T. Jinrui, Y. Xianggen, Z. Zhe, “Modeling Technology in Traveling- Wave Fault Location, “ TELEKOMNIKA, Vol. 11, No. 6, pp. 3333-3340, Jun. 2013.

[79] J. R. Marti “Accurate modeling of frequency-dependent transmission lines in electromagnetic transient simulations, “ IEEE Trans. Power Apparatus and Syst., Vol. PAS-101, No. 1, pp. 147-157, Jan. 1982.

[80] M. Rui, H. Renmu, W. Peng, “A Multiobjective Operation Planning Model With Transmission Constraints and Load Characteristic of Distribution System, “ Transmission and Distribution Conference and Exhibition, 2005/2006 IEEE PES, Dallas, TX, May 2006.

[81] Saber, A. Y., Alam, A. R., ”Short-term load forecasting using multiple linear regression for Big Data< In Computational Intelligence (SSCI), 2017 IEEE Symposium Series on, pp. 1-6.

[82] B. Zhao, J. Cao, Z. Zhu and H. Zhang, "A new transient voltage stability prediction model using Big Data analysis," IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia), Melbourne, VIC, pp. 1065-1069, 2016.

[83] Mao, M., Wang, Y., Yue, Y., & Chang, L. (2017, October), “Multi-time scale forecast for schedulable capacity of EVs based on Big Data and machine learning,” In Energy Conversion Congress and Exposition (ECCE), 2017 IEEE, pp. 1425-1431, IEEE.

[84] Wang, Y., Chen, Q., Kang, C., & Xia, Q. “Clustering of electricity consumption behavior dynamics toward Big Data applications,” IEEE transactions on smart grid, Vol. 7, No. 5, pp. 2437-2447, 2016.

[85] Song, Y., Wang, W., Zhang, Z., Qi, H., & Liu, Y., “Multiple Event Detection and Recognition for Large-Scale Power Systems Through Cluster-Based Sparse Coding,” IEEE Transactions on Power Systems, Vol. 32, No. 6, pp. 4199-4210, 2017.

[86] Dahal, O. P., Brahma, S. M., & Cao, H. (2014) "Comprehensive clustering of disturbance events recorded by phasor measurement units," IEEE Transactions on Power Delivery, Vol. 29, No. 3, pp. 1390-1397, 2014.

[87] S. Ye, X. Wang, and Z. Liu, "Dual-stage feature selection for transient stability assessment based on support vector machine," in Proc. IEEE CSEE, Vol. 30, pp. 28–34, 2010.

[88] Wang, B., Fang, B., Wang, Y., Liu, H., & Liu, Y. "Power system transient stability assessment based on Big Data and the core vector machine," IEEE Transactions on Smart Grid, Vol. 7, No. 5, pp. 2561-2570, 2016.

[89] S. K. Tso, X. P. Gu, Q. Y. Zeng, and K. L. Lo, "An ANN-based multilevel classification approach using decomposed input space for transient stability assessment," Elect. Power Syst. Res., Vol. 46, No. 3, pp. 259–266, 1998.

[90] Zhu, A., Li, X., Mo, Z., & Wu, R. (2017, September) "Wind power prediction based on a convolutional neural network," IEEE International Conference on In Circuits, Devices and Systems (ICCDS), pp. 131-135, 2017

[91] Wang, X. Z., Zhou, J., Huang, Z. L., Bi, X. L., Ge, Z. Q., & Li, L. (2016, March) "A multilevel deep learning method for Big Data analysis and emergency management of power system," IEEE International Conference on Big Data Analysis (ICBDA), pp. 1-5, 2016

[92] J. Hao et al., "Short-Term Power Load Forecasting for Larger Consumer Based on TensorFlow Deep Learning Framework and Clustering-Regression Model," 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, pp. 1-6, 2018.

[93] Stojanovic, J., Jovanovic, M., Gligorijevic, Dj., Obradovic, Z. " Semi-supervised learning for structured regression on partially observed attributed graphs" Proceedings of the 2015 SIAM International Conference on Data Mining (SDM 2015) Vancouver, Canada, April 30 - May 02, 2015.

[94] V. Radosavljevic, S. Vucetic, Z. Obradovic, "Neural Gaussian Conditional Random Fields," Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Nancy, France, September, 2014.

[95] V. Radosavljevic, Z. Obradovic, S. Vucetic, "Continuous Conditional Random Fields for Regression in Remote Sensing, " Proc. 19th European Conf. on Artificial Intelligence, August, Lisbon, Portugal, pp. 809-814, 2010.

[96] N. Djurovic, V. Radosavljevic, Z. Obradovic, S. Vucetic, "Gaussian Conditional Random Fields for Aggregation of Operational Aerosol Retrievals," IEEE Geoscience and Remote Sensing Letters, Vol. 12, No. 4, pp. 761 – 765, 2014.

[97] A. R. Hileman, "Insulation Coordination for Power Systems," CRC Taylor and Francis Group, LLC, 1999.

[98] P. A. Kuntz, R. D. Christie, S. S. Venkata, "A Reliability Centered Optimal Visual Inspection Model for Distribution Feeders," IEEE Transactions on Power Delivery, Vol. 16, No. 4, pp. 718-723, October 2001.

[99] T. Dokic, M. Kezunovic, "Optimized Asset Management in Distribution Systems Based on Predictive Risk Analysis," Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion - MEDPOWER, Dubrovnik, Croatia, November 2018.

[100] D. T. Radmer, P. A. Kuntz, R. D. Christie, S. S. Venkata, R. H. Fletcher, “Predicting Vegetation-Related Failure rates for Overhead Distribution Feeders,” IEEE Transactions on Power Delivery, Vol. 17, No. 4, pp. 1170-1174, October 2002.

[101] R. Sievanen and T. E. Burk, “Adjusting a process-base growth model for varying site conditions through parameter estimation,” Can. J. Forest Res., Vol. 23, No. 9, pp. 1837–1851, Sept. 1993.

[102] D. C. Hamlin and R. A. Leary, “An integro—Differential equation model of tree height growth,” in Proc. IUFRO Conf., vol. 2, Minneapolis, MN, Aug. 1987, pp. 683–690.

[103] B. D. Russell, C. L. Benner, J. Wischkaemper, W. Jewell, J. McCalley, “Reliability Based Vegetation Management Through Intelligent System monitoring,” PSerc Report 07-31, September 2007, [Online] Available: https://pserc.wisc.edu/documents/publications/reports/2007_reports/T-27_Final-Report_Sept-2007.pdf.

[104] J. A. Wischkaemper, C. L. Benner, B. D. Russell, “Electrical Characterization of Vegetation Contacts with Distribution Conductors – Investigation of Progressive Fault Behavior,” Transmission and Distribution Conference and Exposition, Chicago, IL, April 2008.

[105] P.-C. Chen, T. Dokic, N. Stokes, D. W. Goldberg, M. Kezunovic, “Predicting Weather-Associated Impacts in Outage Management Utilizing the GIS Framework,” IEEE/PES Innovative Smart Grid Technologies Latin America (ISGT-LA), Montevideo, Uruguay, October 2015.

[106] P. A. Kuntz, R. D. Christie, S. S. Venkata, “Optimal Vegetation Maintenance Scheduling of Overhead Electric Power Distribution Systems,” IEEE Transactions on Power Delivery, Vol. 17, No. 4, pp. 1164-1169, October 2002.

[107] J. Wingfield, “New York Power Authority Develops Vegetation Management Solution for High-Voltage Transmission Lines with GIS,” ArcNews, Vol. 27, No. 1, pp. 35-, ESRI Spring 2005.

[108] IEEE Standards Association “IEEE Std. C62.82.1-2010 - IEEE Standard for Insulation Coordination –Definitions, Principles, and Rules,” 2010.

[109] Rowland, S. M., and S. Bahadoorsingh. "A Framework Linking Insulation Ageing and Power Network Asset Management." IEEE International Symposium on Electrical Insulation, (ISEI), Vancouver, BC, Canada, 2008.

[110] Tzimas, Antonios, and Simon M. Rowland. "Risk estimation of ageing outdoor composite insulators with Markov models." IET generation, transmission & distribution 6.8 (2012): 803-810.

[111] Tzimas, Antonios, et al. "Asset management frameworks for outdoor composite insulators." IEEE Transactions on Dielectrics and Electrical Insulation, Vol. 19, No. 6, pp. 2044 – 2054, 2012.

[112] A. M. Mousa, K. D. Srivastava, “Effect of shielding by trees on the frequency of lightning strokes to power lines,” IEEE Transaction on Power Delivery, Vol. 3, No. 2, pp. 724-732, April 1988.

[113] A. M. Mousa, “A study of the engineering model of lightning strokes and its application to unshielded transmission lines,” PhD dissertation, The University of British Columbia, Aug. 1986.

[114] T. Sadovic, et al., “Expert System for Transmission Line Lightning Performance Determination”, CIGRE Int. Colloq. on Power Quality and Lightning, Sarajevo, Jun. 2012.

[115] Á. L Orille-Fernández, N. Khalil., and S. B. Rodríguez, “Failure risk prediction using artificial neural networks for lightning surge protection of underground MV cables,” IEEE Transactions on Power Delivery, Vol. 21, No. 3, pp. 1278-1282, 2006.

[116] R. Shariatinasab, et al., "Probabilistic evaluation of optimal location of surge arresters on EHV and UHV networks due to switching and lightning surges." IEEE Transactions on Power Delivery, Vol. 24, No. 4, pp. 1903-1911, 2009.

[117] A. P. Sakis Meliopoulos, „Lightning and Overvoltage Protection,“ In: D. G. Fink, H. W. Beaty, editors. Standard handbook of electrical engineering. 16 th ed. New York: McGraw-Hill; 2007, ISBN: 9780071762311 .

[118] IEEE Standards, “International Standard IEC 62539 – IEEE 930: Guide for the statistical analysis of electrical insulation breakdown data,” 2007

[119] 3D Land Mapping, Jet Propulsion Laboratory, California Institute of Technology. [Online] Available: http://lidarradar.jpl.nasa.gov/

[120] M. Kezunovic, L. Xie, and S. Grijalva, "The role of Big Data in improving power system operation and protection," 2013 Bulk Power Systems Dynamics and Control Symposium – IREP’2013, pp. 1-9, Rethymno, Greece, Aug. 2013.

[121] ArcGIS, Esri. [Online] Available: https://www.arcgis.com

[122] Python, [Online] Available: https://www.python.org/

[123] David J. Buckey, Bgis Introduction to GIS, 11/10/2013. [Online]. Available: http://bgis.sanbi.org/gis-primer/

[124] GARMIN, “What is GPS?” [Online] Available: http://www8.garmin.com/aboutGPS/

[125] National Geospatial-Intelligence Agency, “NGA Standardization Document – Time-Space-Position Information (TSPI), NGA.STND.0019_2.0 2012-04-05.

[126] Network Time Foundation, “NTP: The Network Time Protocol,” [Online] Available: http://www.ntp.org/

[127] IEEE Standards, “IEEE 1588-2002,” 8 November 2002.

[128] A. von Meier, A. McEachern, “Micro-synchrophasors: a promising new measurement technology for the AC grid,” i4Energy Seminar October 19, 2012.

[129] V. Radosavljevic, K. Ristovski, Z. Obradovic, “Gaussian Conditional Random Fields for Modeling Patients’ Response to Acute Inflammation Treatment,” Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013.

[130] B. M. Ayyub, “Risk Analysis in Engineering and Economics,” CRC Press, 2003, ISBN: 978-1-4665-1826-1, pp. 57-60.

[131] M. Kezunovic, Z. Obradovic, T. Dokic, B. Zhang, J. Stojanovic, P. Dehghanian, and P. -C. Chen,"Predicting Spatiotemporal Impacts of Weather on Power Systems using Big Data Science,"Pedrycz, Witold, Chen, Shyi-Ming (Eds.), Springer Verlag, Data Science and Big Data: An Environment of Computational Intelligence, ISBN 978-3-319-53474-9, 2017.

[132] IEEE Standards, “IEEE 1588-2002,” 8 November 2002.

[133] Qin, T., Liu, T. Y., Zhang, X. D., Wang, D. S., & Li, H., “Global ranking using continuous conditional random fields,” In Advances in neural information processing systems, pp. 1281-1288, 2009.

[134] V. Radosavljevic, S. Vucetic, Z. Obradovic, "Continuous Conditional Random Fields for Regression in Remote Sensing," Proc. 19th European Conf. on Artificial Intelligence, August, Lisbon, Portugal, 2010.

[135] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li, "Global Ranking Using Continuous Conditional Random Fields," in Proceedings of NIPS'08, 2008, Vol. 21, pp. 1281–1288.

[136] W. Jewell, J. Warner, J. McCalley, Y. Li, S. R. K. Yeddanapudi, "Risk-Based Resource Allocation for Distribution System Maintenance," PSerc Report 06-26, August 2006, [Online] Available: https://pserc.wisc.edu/documents/publications/reports/2006_reports/T-24_Final-Report_Aug-2006.pdf.

[137] ArcGIS, "Iso Cluster Unsupervised Classification," [Online] Available: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/iso-cluster-unsupervised-classification.htm

[138] W. Tian, Y. Zheng, R. Yang, S. Ji, and J. Wang, "A Survey on Clustering based Meteorological Data Mining," International Journal of Grid Distribution Computing, Vol. 7, No. 6, pp. 229-240, 2014.

[139] M. Kezunovic, T. Dokic, R. Said, "Optimal Placement of Line Surge Arresters based on Predictive Risk Framework using Spatio-Temporally Correlated Big Data," CIGRE Paris, August 2018.