

# Predictive and Prescriptive Analytics for Managing the Impact of Hazards on Power Systems

by

Elnaz Kabir

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Industrial and Operations Engineering)  
in The University of Michigan  
2021

Doctoral Committee:

Professor Seth D. Guikema, Chair  
Professor Brian T. Denton  
Assistant Professor Johann Gagnon-Bartsch  
Professor Steven M. Quiring

Elnaz Kabir  
ekabir@umich.edu  
ORCID iD: 0000-0002-5211-7433

© Elnaz Kabir 2021

## **DEDICATION**

To my parents and my spouse,  
for their unconditional love and support.

## ACKNOWLEDGEMENTS

I am profoundly grateful for the many people who have supported me throughout my PhD journey. Special thanks to my adviser and mentor, Seth Guikema. I have learned so much from you through the years, you have taught me how to learn, how to teach, how to be a good colleague and collaborator. Thank you for your encouragements, patience, and support.

I would also like to thank the remaining members of my committee for their input and guidance, and the many productive conversations and collaborations. Brian Denton, thank you for letting me teach at the IOE and always being supportive to me; Steven Quiring, thank you for your guidance and help throughout my studies, and for providing a realistic setting for my research; Johann Gagnon-Bartsch, thank you for teaching me the foundations of statistical learning at the beginning of my Ph.D. journey.

I am especially grateful to my parents, Zahra and Ahmad, who have raised me to be the person I am today. I am forever thankful to you for your unwavering support throughout the years, and instilling in me a passion for learning and pursuing my goals. I wish you were here, and I could share this achievement with you. I would also like to thank my wonderful in-laws, Tooba and Ebrahim, for the kindness and continual support they provided throughout these years.

A very special thank you to my spouse, Esmail, for his love, unconditional support and all his positive thoughts. Without your love and support, I would have never come this far. Thank you for being there for me every step of the way and giving me the endurance and strength to arrive at this milestone. Words cannot express my gratitude for having you.

I owe gratitude to my wonderful and inspiring siblings for always being there for me since day one and despite the distance. Thanks for helping me learn from you, and being my best friends. Thank you for loving me at my worst, just as much as my best.

My thanks goes out to my wonderful friend, Sepideh. Thank you for always being there for me, and listening to me when I needed to talk, and talking when I needed to listen. Thank you to the many people in Ann Arbor who have made this place home. I have made some of the best friends I could ever ask for - Delaram, Mohammad, Sajedah, Mohsen, Niloofar, Mehrdad, Neda, Amin, Mahnaz and Iman. Thank you for creating many fond memories

that I will treasure. Without you, I would not have been able to enjoy my life here in Ann Arbor throughout these years. Friday night gatherings have been a highlight of my weeks.

Thank you to my classmates, friends in the department, and especially the Guikema research group - Tom Logan, Thomas Chen, Tim William, Chengwei Zhai, Anna White, Julia Coxen, and Tessa Swanson. You have encouraged me throughout the years, and you made me look forward to coming into the office every day.

# TABLE OF CONTENTS

DEDICATION . . . . .	ii
ACKNOWLEDGEMENTS . . . . .	iii
LIST OF FIGURES . . . . .	viii
LIST OF TABLES . . . . .	xi
ABSTRACT . . . . .	xiii
<b>CHAPTER</b>	
<b>I. Introduction . . . . .</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Background on Storm Outage Management . . . . .	1
1.3 Use of Analytics in Storm Outage Management . . . . .	4
1.4 Dissertation Outline . . . . .	5
<b>II. Predicting Thunderstorm-Induced Power Outages To Support Util-     ity Restoration . . . . .</b>	<b>8</b>
2.1 Introduction . . . . .	8
2.1.1 Research Motivation . . . . .	8
2.1.2 My Contributions to the Literature . . . . .	9
2.1.3 Chapter Organization . . . . .	11
2.2 Literature Review . . . . .	11
2.2.1 Power Outage Prediction . . . . .	11
2.2.2 Learning from Zero-inflated or Imbalanced Data . . . . .	14
2.3 Data Description and Methods . . . . .	15
2.3.1 Data Description . . . . .	15
2.3.2 Models Implemented . . . . .	18
2.4 Computational results and analysis . . . . .	21
2.4.1 Probabilistic and Point Predictions . . . . .	22
2.4.2 Partial Dependence Plot and Variable Importance . . . . .	26

2.4.3	Prediction for New Storms . . . . .	28
2.5	Summary and Conclusions . . . . .	29
<b>III. Adaptive Two-stage Bayesian Model Averaging for Estimating the Impact of Hazards on Power System Service . . . . .</b>		<b>31</b>
3.1	Introduction . . . . .	31
3.1.1	Research Motivation . . . . .	31
3.1.2	My Contributions to the Literature . . . . .	33
3.1.3	Chapter Organization . . . . .	34
3.2	Literature Review: Weather-induced Outage Prediction Modeling . .	34
3.3	Model . . . . .	37
3.3.1	Model Averaging . . . . .	37
3.3.2	Proposed model: Adaptive two-stage BMA . . . . .	38
3.4	Case Study: Predicting The Number Of Customers Interrupted . . .	44
3.4.1	Data Description . . . . .	44
3.4.2	Evaluation metric . . . . .	46
3.4.3	Computational results and analysis . . . . .	46
3.5	Summary and Conclusions . . . . .	50
<b>IV. An Assessment of Drivers of Power System Damage During Severe Weather . . . . .</b>		<b>52</b>
4.1	Introduction . . . . .	52
4.1.1	Research Motivation . . . . .	52
4.1.2	My Contributions to the Literature . . . . .	53
4.1.3	Chapter Organization . . . . .	54
4.2	Literature Review . . . . .	54
4.2.1	Modeling Power System Damage . . . . .	54
4.2.2	Bayesian Belief Network . . . . .	55
4.3	Data Description . . . . .	60
4.4	Method and Computational Results . . . . .	61
4.4.1	Network structure learning . . . . .	61
4.4.2	Belief Propagation Analysis . . . . .	63
4.5	Discussion and Insights . . . . .	70
4.6	Summary and Conclusions . . . . .	71
<b>V. A Multi-stage Stochastic Crew Coordination Model for Power Outage Restoration . . . . .</b>		<b>73</b>
5.1	Introduction . . . . .	73
5.2	Literature Review . . . . .	75
5.3	Problem Statement . . . . .	77
5.4	Mathematical formulation . . . . .	83
5.4.1	Multi-stage Stochastic Mixed-Integer Program Model . . .	83

5.4.2	Linear MSSP Crew Coordination Model . . . . .	88
5.5	Computational results and analysis . . . . .	90
5.5.1	Assessing the performance of the MSSP . . . . .	90
5.5.2	Application of the proposed stochastic model . . . . .	92
5.6	Summary and Conclusions . . . . .	95
<b>VI. Conclusions and Future Research . . . . .</b>		<b>96</b>
6.1	Summary and Conclusions . . . . .	96
6.2	Future Research . . . . .	99
<b>BIBLIOGRAPHY . . . . .</b>		<b>101</b>

## LIST OF FIGURES

### Figure

1.1	Four steps performed in a storm outage management. . . . .	2
1.2	Three phases of analytics. . . . .	4
2.1	The red area in the left figure shows the total coverage area of my data set. The right figure plots the number of customers vs number of customers with power outages for all the instances. It shows the concentration of customer outages around zero, and zero-inflation issue in the data. . . . .	16
2.2	The illustration of the proposed two-stage framework. In the preprocessing step, either training data are rebalanced or misclassification costs of the non-zero class instances are enhanced in the classifier. In the first stage, a classifier predicts the probability of each instance, $x_i$ , having non-zero outages, $P_{x_i}$ . In the second stage, firstly, a simulator generates random numbers between 0 and 1, $u_i, \forall i = 1, \dots, S$ ; then, $QRF^0$ and $QRF^1$ predict two separate distributions for each instance by assuming that the instance belongs to the zero or non-zero class respectively. Finally, by considering the values of $P_{x_i}$ and $u_i$ s, random records are generated from the predicted distributions. These records together estimate the full probability distribution for each instance $x_i$ . . . . .	18
2.3	Comparison of the boundaries obtained by SVDDs with linear and nonlinear kernels. . . . .	21
2.4	Continuous Ranked Probability Score (CRPS) for three typical grid cells with observed outages equal to (a) 0, (b) 104, and (c) 1470 for the storm occurred in April 2006. The green solid curve shows the predicted CDF, and the black vertical line shows the actual value for the number of customers without power for each grid cell examples. The shaded green area under the green dashed curve is the CRPS for the predicted CDF. CRPS is a measure for presenting the variation of the predicted CDF from the actual value. The closer the green solid curve is to the black stepwise function, the more accurate is the predicted CDF and smaller is the CRPS value. . . . .	22

2.5	Comparing probabilistic predictions of all models based on CRPS . . . . .	23
2.6	Comparison of the point estimates obtained by different models in terms of the MAE. The overall MAEs as well as the MAEs of the records with zero outages and records with non-zero outages are calculated separately. . . . .	24
2.7	Analyzing the accuracy of cost-sensitive RF-QRF model for difference values of the cost parameter. . . . .	25
2.8	Partial dependence plot for the 8 most important variables . . . . .	27
2.9	Comparison between spatial distribution of actual outages and estimated ones for the storm occurred in April 2006 . . . . .	29
3.1	This figure summarizes my proposed adaptive two-stage Bayesian model averaging algorithm. There are $K$ base learners represented by $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K$ . Each base learner independently predicts probability distribution of response variable $y$ for each covariate vector $x$ knowing the cluster this record belongs to, $C(x)$ . The final probabilistic prediction for covariate vector $x$ is a weighted average over all predicted probability distributions. The weights of models in the final prediction is a multinomial logit model of covariate $x$ . Parameters of the multinomial logit model as well as all the base learners are updating over time using newly observed data. . . . .	40
3.2	Spatial extent of the service territory for AEP Ohio, I&M Power and AEP Texas. The total service territory is divided into 46 subareas that are defined by the utility company and shown in different colors. . . . .	44
3.3	CH factor is estimated for different number of clusters. Having six clusters results in the optimum CH value. . . . .	45
3.4	Data is clustered based on the static variables, and different colors show the resultant six clusters. . . . .	45
3.5	Decrease of the CRPS error of the BMA model as a function of iterations. . . . .	46
3.6	Boxplot comparison of the CRPS errors of the proposed BMA model with its base learners. . . . .	47
3.7	Comparison of the MAE error of the proposed BMA model with its base learners. . . . .	48
4.1	Spatial extent of the service territory for six operating companies serving 29 districts in 10 U.S. states. The total service territory is divided into 29 subareas that are defined by the AEP utility company and shown in different colors. . . . .	58

4.2	The Bayesian belief network representing the associations among meteorological variables (shown by white circles), damaged assets (shown by yellow circles) and customer interruptions (shown by brown circle) in the power grid.	62
4.3	Results of the ANOVA test for each variable and damage type. The boxplots show the distribution of damages in various levels of each variable. The connected red points indicate the mean damage in each group. P-value of the ANOVA test, the maximum and minimum damage means are also shown in each sub-figure. . . . .	64
4.4	The daily proportion of different damage types in the service territory of each operating company. . . . .	65
4.5	The distributions of different variables in the service territory of each operating company. . . . .	67
4.6	The comparison of the MI between different meteorological variables and damage/customer interruptions variables. The larger the size of the circle is, the larger the MI exists between the two corresponding variables. . . . .	69
5.1	The crew supply network structure for power restoration in disasters. This figure demonstrates my network including utility companies, contracting agencies, staging areas, stores and demand points. It also shows resource flows between these entities for my repair crew coordination problem. . . . .	79
5.2	Decision-making process in my optimization model include two types of variables: (i) here and now decisions and (ii) wait and see decisions. . . . .	80
5.3	A scenario tree example for a MSSP with four stages and eight scenarios (left figure) and the decomposed scenario tree to represent the non-anticipativity constraints (right figure). . . . .	83
5.4	Sensitivity of the model to parameter $a$ . . . . .	94
5.5	Sensitivity of the model to parameter $Ed$ . . . . .	94
5.6	Sensitivity of the model to increase in demand. . . . .	95
5.7	Sensitivity of the model to the utility function $f(\cdot)$ . . . . .	95

## LIST OF TABLES

**Table**

2.1	Literature review table for storm power outage predictive modeling . . . . .	13
2.2	Comparing the accuracy of various models . . . . .	25
2.3	Variable importance in the final cost-sensitive-RF-QRF model . . . . .	28
3.1	Average coefficients of the multinomial logit model in the AT-BMA algorithm are shown in this table. Each row of the table indicates the average of coefficient values for each model. The multinomial logit model is built of 21 variables including an intercept, 15 continuous weather-related variables, and 5 binary variables associated with the cluster covariate. The larger each variable coefficient is, the better prediction the model makes for records with larger values of that variable. . . . .	49
4.1	List of explanatory variables used for learning the model . . . . .	59
4.2	List of all variables with their ranges. Each continuous variable is transformed into a categorical one with at most five levels. Each column of the table indicates the range of values falling in the corresponding category (level). . . . .	61
4.3	List of the features I study their effects on the damage and customers interrupted variables using forward BP. For each studied feature, I also show its confounders that have backdoor path to the damage variables (i.e., variables that influence both the studied feature and the damage or customers interrupted variables without being an intermediate cause in the causal pathway between the studied feature and the damage or customers interrupted variables). The confounders are discovered using the BBN shown in Figure 4.2. . . . .	63
4.4	Summary of the backward belief propagation results. In each analysis, I set a damage/outage variable equal to its highest level and compute the generalized Bayes factor (GBF) for every combination of the set of target variables. I then report the condition under which the GBF value is maximized. In some cases, more than one condition may result in maximum GBF. . . . .	68

4.5	Ranking of the variables based on their MI with the damage data. Lower numbers represents higher MI between the two corresponding variables. . .	70
5.1	The description of indices, parameters and decisions of the model. . . . .	84
5.2	Results from solving test problems with Gurobi solver and estimated values of the multi-stage stochastic program. . . . .	92
5.3	Characteristics of parameters for the test problems. . . . .	93
5.4	Approximate crew hour need for fixing power system damages. . . . .	93

## ABSTRACT

Natural hazards and extreme weather events have the potential to cause significant disruptions to the electric power grid. The resulting damages are, in some cases, very expensive and time-consuming to repair and they lead to substantial burdens on both utilities and customers. The frequency of such events has also been increasing over the last 30 years and several studies show that both the number and intensity of severe weather events will increase due to global warming and climate change. An important part of managing weather-induced power outages is being properly prepared for them, and this is tied in with broader goals of enhancing power system resilience. Inspired by these challenges, this thesis focuses on developing data-driven frameworks under uncertainty for predictive and prescriptive analytics in order to address the resiliency challenges of power systems. In particular, the primary aims of this dissertation are to:

1. Develop a series of predictive models that can accurately estimate the probability distribution of power outages in advance of a storm.
2. Develop a crew coordination planning model to allocate repair crews to areas affected by hazards in response to the uncertain predicted outages.

The first chapter introduces storm outage management and explains the main objectives of this thesis in detail. In the second chapter, I develop a novel two-stage predictive modeling framework to overcome the zero-inflation issue that is seen in most outage related data. The proposed model accurately estimates customer interruptions in terms of probability distributions to better address inherent stochasticity in predictions. In the next chapter, I develop a new adaptive statistical learning approach based on Bayesian model averaging to formulate model uncertainty and develop a model that is able to adapt to changing conditions and data over time. The fourth chapter uses Bayesian belief network to model the stochastic interconnection between various meteorological factors and physical damage to different power system assets. Finally, in chapter five, I develop a new multi-stage stochastic program model to allocate and relocate repair crews in impacted areas during an extreme weather event to restore power as quickly as possible with minimum costs.

This research was conducted in collaboration with multiple power utility companies, and some of the models and algorithms developed in this thesis are already implemented in those companies and utilized by their employees. Based on actual data from these companies, I provide evidence that significant improvements have been achieved by my models.

# CHAPTER I

## Introduction

### 1.1 Motivation

Every year, millions of customers lose their power in the U.S. because of weather related events, and in some areas they are left without power during difficult conditions such as extreme cold or extreme heat. We cannot avoid many of these outages; however, by being better prepared for the event, we can reduce losses and restore power more quickly. Power outage forecasts can help utility companies and emergency officials make better decisions in terms of budget allocation and resource planning.

Power infrastructure is a complex system with many constraints and much inherent uncertainty. To appreciate the complexities and nuances of this system, I start by describing the general background on storm outage management, including different types of practices that can be done to improve the resilience of this critical system. Next, I give a chapter-by-chapter summary of the remainder of this thesis.

### 1.2 Background on Storm Outage Management

Natural hazards such as severe weather events including hurricanes, thunderstorms, winter storms, lightning storms, and tropical cyclones have the potential to damage electric utility transmission and distribution systems and result in long-term and widespread loss of electrical power for affected locations [39, 83]. It is reported that 87% of the total power outages in the U.S. are caused by severe weather events. Loss of electrical power can be either from direct impact to a power generation or distribution system, or indirectly from other objects like fallen trees and branches, which are knocked into overhead lines and cause the poles to snap and dislodge overhead lines from crossarms [70, 51]. This type of damage will disrupt electrical service until the physical facilities can be replaced or repaired. Large storms can result in a massive number of electric power outages, sometimes taking from days



Figure 1.1: Four steps performed in a storm outage management.

to weeks to repair. Due to the complex interdependencies that exist between the electric infrastructure and other critical lifelines in the U.S., electric power outages can adversely affect national security, digital economy, public health, and the environment, which results in huge losses. A congressional research service study done in 2012 estimates the annual inflation-adjusted cost of weather-related outages at 25 to 55 billion [9].

To restore service after a major storm, utilities send a large number of crews into the affected areas. The number of dispatched maintenance crews depends on the scope of the damage and the number of impacted customers. The lowest level is the normal day-to-day outages due to minor storms, animal contact, broken tree limbs, etc., which can be handled locally through conventional outage management system processes. However, in a larger scale such as outages caused by severe storms impacting multiple operating areas, the internal crews might not be enough and so, a centralized storm outage management is required to thoroughly coordinate both internal crews and extra resources called in from other utilities or contracting agencies [70]. Storm outage management is then utilized when large-scale storms cause massive amounts of physical damage, and requires a large number of repair crews for restoring customer service [69]. Figure 1 shows the four steps performed in a storm outage management [69].

**Damage prediction:** The first stage in storm outage management is damage prediction. This involves using weather forecasts, asset information and historical data to predict the amount of damage a storm will produce, the number of people without power, the resources required for restoration and the resulting time needed to restore service to customers. This is a rough estimate, but it allows the utility companies to get resources into place or on stand-by, speeding up the restoration process after the storm hits. A diverse collection of engineering and statistical models are currently used to estimate the geographical distribution of power outage probabilities stemming from the storms to aid in the preparedness and recovery efforts [70, 69]. In these models, power outage can be reported in two different units: (1) population without power (customer outages), and (2) number of outages. If something (e.g., a tree falls on a line) causes physical damage to the electric power system during a storm, the closest protective device upstream is activated to handle this damage [113]. All customers on the isolated portion of the system lose power. In this situation, such a scenario is considered to be a single outage (i.e., activation of a protective device caused by physical damage requiring repair by a crew). One outage, hence, may be associated with a few or a

large number of physical damage, and from one to many customers losing power [113].

**Repair crew and material staging:** The repair crew and material staging process involves specifying the number of crews and materials required for restoring the affected areas in an appropriate amount of time, and if more resources are needed, to make immediate arrangements for external crews to be called in. It also involves making decisions about where to stage these crews and materials and where they should be lodged and fed. Staging is a logistically challenging task because it includes making arrangements for housing and feeding a large crew. The crew and material staging function would take the damage prediction results and make initial assignments for internal crews, identify need and locations for external crews and determine need and locations for materials ahead of time [113].

**Damage assessment:** Damage assessment process starts once the storm has hit. It involves filing reports based on customer calls to provide service outage information, sending trained teams of crews to investigate the type and location of damages, or using some automated metering tools to detect outages [113, 105, 13]. Using the collected damage information, managers can then decide how to dispatch their repair crews and materials that are staged before the storm. Using the verified damage information, and number of assigned crews and materials to each location, an estimate for customers as to how long they will be without power can be computed. In this stage, predicted damage estimates and the customer outages obtained from previous phases are converted to the verified damage information and customer outages [69].

**Restoration management:** The restoration process, which is done after the storm passes through the utility's service region, lasts the longest. In this process, repair crews are initially dispatched to areas according to storm damage and are allocated to work specific substations and feeders to restore service. Crews can also be reallocated as necessary by the storm outage coordinator. It is a challenging task to track crew progress in real time manner because there are many crews operating at the same time coming in from different utilities or contracting firms. Given the number of crews allocated per feeder, time to restore power to each customer is estimated. This is based on predefined guidelines for how restoration is to take place. In the restoration process, repair tasks are prioritized such that fixing the damages that return power to larger number of customers has higher priority than other tasks. In this stage, total cost estimates for making the repairs based on the crew allocation can also be made.

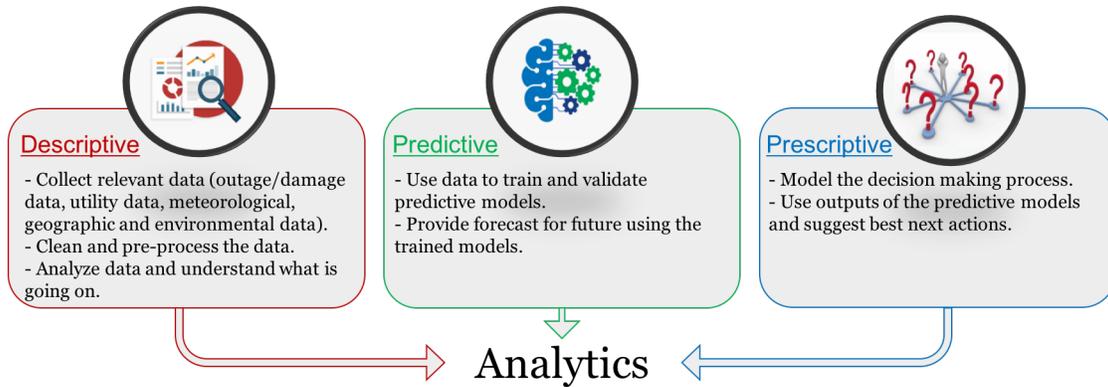


Figure 1.2: Three phases of analytics.

### 1.3 Use of Analytics in Storm Outage Management

The main goal of this thesis is to use analytics in order to do better storm outage management and address the needs of modern power systems. Analytics start with data and once we have data, we need to analyze it to have better understanding of the system. Building on these data, we deploy statistical and machine learning models in order to forecast future scenarios. Finally, we use math and optimization to drive decision making. Figure 1.2 represents these three steps of analytics.

In the context of storm outage management, I start with data collection and data pre-processing. I use utility provided outage data. There are some other asset data that are utility specific and I have to get them from the utility companies. I also obtain other data such as weather forecasts, geographic and environmental data from related sources. In each chapter, I provide a detailed explanation about the source of the data used for that study. The next step is to clean the data and prepare it for analysis. The better data I collect and prepare in this stage, the better predictive and prescriptive models I can obtain.

In the predictive modeling phase, I train, test and validate statistical and machine learning models. Once these models are developed, real-time outage predictions can be made to predict the impacts of a hazard on the power system. As more data is collated, data and models can be updated as well. Outage predictive models transfer data into informed decision making. The more accurate these estimates are; the better preparation decisions can be made by utility companies. This helps utilities better plan their resource needs, and increase the rate of restoration. Finally, before the weather event, especially extreme ones, utility companies start planning based on the predictions. Descriptive and predictive analytics are done in advance of a storm or hazard. However, prescriptive analytics starts prior to the event but continues until the end of restoration process. Chapter II and III

are focused on predictive analytics, and Chapter V is concentrated on prescriptive analytics for power systems. Descriptive analytics is an important part of each chapter while it is the main focus of Chapter IV.

## 1.4 Dissertation Outline

This dissertation is structured as follows. Chapters II-V are based on independent academic papers. This first chapter introduces the broader area of power outage management and how predictive and prescriptive analytics can be utilized in this context. In the following I summarize each chapter.

**Chapter II - Predicting Thunderstorm-Induced Power Outages To Support Utility Restoration:** Extreme weather events such as hurricanes and thunderstorms have substantial impacts on power systems, posing risks and inconveniences due to power outages. Developing models predicting outage variables (e.g., number of customers who experienced an outage, outage duration and number of physical damages) prior to a storm facilitates disaster response decision-making by electric power utilities as well as other organizations of critical importance to society. Typically, the area of interest is divided into grid cells and the number of outages is forecasted for each grid cell. Developing models based on real-world infrastructure data in resolutions smaller than census tract or county level is a challenge due to the zero-inflation or imbalance in the data. This occurs whenever there exist significantly more observations of zero outages than non-zero. This issue leads to bias and inaccuracy in predictive modeling. In addition, power outages are stochastic and there always exists irreducible variability in outage predictions. However, developing models estimating power outages with a single value gives the decision maker a false impression of perfect accuracy.

Inspired by these challenges, in Chapter II, I develop a novel two-stage predictive modeling framework to overcome the zero-inflation issue and accurately estimate outages in terms of probability distributions to better address inherent stochasticity in predictions. It integrates mixture models with imbalanced-learning techniques. Validating my approach using actual thunderstorm data, I demonstrate that it captures the effects of climatological, geographical, and environmental variables on the power systems and offers more accurate point and probabilistic predictions compared to existing modeling approaches. This modeling framework is currently being implemented by a large utility company in the central Gulf Coast region, and they used for predicting the number of thunderstorm-induced customers without power.

**Chapter III - Adaptive Two-stage Bayesian Model Averaging for Estimating the Impact of Hazards on Power System Service:** In Chapter II, I capture uncertain-

ties existing in model outputs and represent them using probabilistic predictions. However, there is another important source of uncertainty in statistical machine learning models called model uncertainty. Due to noise in observations and incomplete coverage of data, selecting one model as the one describing the process is not always a proper approach. This hides the existing uncertainty in the model and results in decision-making that is not well-informed. Furthermore, the power system, climatological, and environmental variables are changing over time. This necessitates models that are able to adapt to changing conditions and data over time, which allows utility companies achieve better outage predictions while investing less time, effort and resources.

Motivated by these research gaps, I develop a new adaptive statistical learning approach based on Bayesian Model Averaging (BMA). Instead of developing one single model, this algorithm is built upon a number of competing base learners. Unlike the classical BMA, I consider a decision-theoretic approach and formulate weights of the base learners with an online multinomial logistic model. This allows the algorithm to assign weights to the base learners that are specific to each newly observed data point according to its features. By using (i) Bayes theorem, (ii) Laplace approximation, and (iii) stochastic gradient ascent, posterior distributions of the parameters of the multinomial logistic model are estimated and updated. Validating my algorithm using daily customer interruption data, I showed that my algorithm results in a more accurate probabilistic prediction than the base learners individually, and yields more accurate predictions as more data are observed. This algorithm is already implemented in the American Electric Power (AEP) company to make daily predictions for the number of customers without power. Although my work is motivated by power system application, my methodology and insights can be implemented in other predictive modeling problems dealing with high model uncertainty.

**Chapter IV - An Assessment of Drivers of Power System Damage During Severe Weather:** Due to the ease of collecting outage variables through an automated system, existing research has focused mostly on modeling the number of outages, number of customers without power, and power outage duration. However, outage focused predictive modeling is not very applicable for making system reinforcement decisions at the asset level.

In this chapter, I study the impacts of meteorological variables on the failure of utility assets including conductors, transformers, and poles. I develop a Bayesian belief network to model the stochastic interconnection between various meteorological factors and physical damage to different power system assets. Hypothesis tests, matching for controlling confounders' effects, maximum relevant explanation, and mutual information are the tools I use to perform belief propagation and variable importance analysis. These techniques help the policy maker (i) understand the effects of each individual variable on the power system

damages, (ii) find the weather conditions that derive the maximum level of damages, and (iii) rank the meteorological factors based on their influence on the power system damages. Using real data of daily damage occurring in districts served by multiple utility companies, I provided them with several critical insights on how to find the vulnerable components of power systems, understand the factors driving outages, and suggest actionable strategies to perform cost-benefit analysis and effective system reinforcement.

**Chapter V - A Multi-stage Stochastic Crew Coordination Model for Power Outage Restoration:** Before an extreme weather event, based on the expected damage, repair crews are often dispatched to impacted areas to be able to start restoration process as quickly as possible. In most large-scale outages, utilities must request crews from other companies. The coordination of crews between different districts in real time is a challenging task, because there are many crews from multiple utilities operating simultaneously. Inspired by this challenge, I develop a new data-driven multi-stage stochastic program (MSSP) methodology for allocation and relocation of repair crews in impacted areas during an extreme weather event to restore power as fast as possible with minimum costs. Due to the inherent uncertainty in damage rates and restoration time, there is a huge uncertainty in demand for which I build a finite set of scenarios, described in the form of a scenario tree. This decision-making framework integrates a MSSP optimization model with a crew demand prediction model. The main feature of this framework is that its decisions are implementable in real time, because these decisions can be adjusted progressively based on realized uncertainty. Numerical results demonstrate the significance of my model. Finally, several key managerial and practical insights in terms of resource allocation are highlighted.

**Chapter VI - Conclusions and Future Research:** The works presented in Chapters II-V make contributions into three important parts of storm outage management including damage prediction, crew staging and restoration management. In Chapter VI, I summarize some of the most important contributions. I also highlight areas of future research that could expand on this work.

## CHAPTER II

# Predicting Thunderstorm-Induced Power Outages To Support Utility Restoration <sup>1</sup>

## 2.1 Introduction

### 2.1.1 Research Motivation

The electric power system is one of the backbones of modern society and economies. Loss of electricity causes considerable inconvenience for residents and widespread economic and non-economic losses. Extreme weather events are the major cause of damage to electric distribution networks and resultant power outages in the U.S. [44]. The U.S. Department of Energy has estimated that annual economic losses due to weather-related power outages are \$25 billion [71]. A key component in reducing losses from weather-induced power outages is being able to predict outages in advance. This helps utilities better plan resource needs, increasing the rate of restoration. This ties in with broader goals of enhancing power system resilience.

In recent decades, the problem of power system resilience has been studied from different perspectives and considerable progress has been made (e.g., [58, 89, 116, 2]). In addressing resilience to extreme weather, models and strategies are exploited in three stages: (i) prior to the event (e.g., [136, 66, 127, 2]), (ii) during the event (e.g., [120, 116]), and (iii) after the event (e.g., [12, 107]). Predicting power outages, hardening existing distribution lines, vegetation management, deploying resources such as back-up distributed generators and automatic tie switches are effective resiliency strategies that can be done prior to the storm with vary degrees of required lead time [89]. Physically changing power systems, controlling power flow in distribution networks, islanding, and self-healing schemes are some of the resiliency activities that can be implemented during a storm. System status evaluation, establishing a

---

<sup>1</sup>Kabir, E., Guikema, S.D., and Quiring, S.M. (2019), Predicting thunderstorm-induced power outages to support utility restoration, IEEE Transactions on Power Systems, 34(6), 4370-4381.

strong bulk power network, and fast load restoration are some practices performed after the storm [116].

However, having a predictive model accurately estimating power outages is a critical part of pre-storm resiliency practices, and it also directly influences the decisions made in all above-mentioned three stages of the power resilience problem [116]. This key observation is also highlighted by the recent survey of Wang et al., [119] on the resilience of power systems. They mention two promising directions for future power outage prediction models: (i) enhancing the accuracy of the predictions by developing new statistical models, and (ii) establishing models that link prediction and hardening investment guidance. This chapter is motivated by this research gap, and provides a comprehensive study to develop models accurately predicting power outages and improve the resilience of power systems.

### 2.1.2 My Contributions to the Literature

In this chapter, a probabilistic modeling approach is developed for the power outage prediction problem, which focuses on outage prediction for thunderstorms, an under-studied type of outage cause. Furthermore, this proposed modeling framework improves the accuracy of the predictions by using statistical techniques to overcome the challenges caused by the zero-inflation property of outage data. The algorithm is trained on data obtained from 11 strong thunderstorms that occurred in Alabama over the past ten years. I seek to address (i) how accurate are these models in providing point estimates? and (ii) how efficient are these models in estimating probability distributions for zero-inflated outages? In the following, I detail the major departures of this chapter from the existing literature and also highlight my high-level approaches and techniques.

**(1) Overcoming the zero-inflation:** Machine learning (ML) models used in the literature have worked well in predicting power outages (e.g., [58, 82]). However, these ML models are based on some assumptions that are often violated in practice. One of these critical violations is that power outage data are often highly zero-inflated. That is, the number of zero outages is significantly higher than the number of non-zero outages. With zero-inflated data, classical ML models struggle to appropriately model the data and make accurate predictions. Ignoring zero-inflation has substantial consequences. First, the estimated parameters and standard errors may be biased toward zeros if parametric regression models are used. Second, the excessive number of zeros can cause over-dispersion. Third, the use of global performance metrics such as overall accuracy induces a bias toward zero; that is, even though the model does not make accurate predictions for the non-zero observations, it has a high overall “accuracy” due to precise predictions made for zero observations. Putting this differently, models can make “accurate” predictions for zero-inflated data by

always predicting zero if global error metrics MAE or RMSE are used, but these predictions are useless in practice because they under-estimate outages in impacted locations.

The two-stage modeling approach used in [37] and [74] is recognized to be a successful approach to tackle both zero-inflation and the complexities existing in the power outage data. Building on their approach, I develop a novel two-stage modeling framework for predicting outages due to thunderstorms. The first stage of my framework is based on random forest, boosting trees and support vector machine classifiers, and determines whether a grid cell (a record in the data set) has at least one customer without power during the storm. To overcome the above-mentioned challenges arising from the zero-inflation, I then incorporate two unbalanced learning techniques, including resampling and cost-sensitive learning into the first stage of this two-stage model. The results show that both techniques, by reducing the bias of the first-stage model towards zero-class data, enhance the accuracy of predictions specifically for the non-zero class data which results in improvement in the overall accuracy. Even though both unbalanced learning techniques improve the point estimate of power outages compared to the traditional two-stage model developed by [74], only the cost-sensitive learning enhances the accuracy of probabilistic predictions significantly. The authors believe that this is because unlike resampling, cost-sensitive learning does not change the distribution of the data.

**(2) Producing probabilistic power outage predictions:** Weather-induced power outages have inherent stochasticity, and uncertainty exists in any outage prediction. However, the power outage prediction literature lacks models accurately estimating the probability distributions of outages [119]. Almost all developed models estimate power outages by a single value (e.g., [58, 74]) rather than a probability distribution, which gives the decision maker a false impression of precision and hides the existing uncertainty.

In the proposed novel two-stage model, the second stage is based on a Quantile Regression Forest (QRF). This allows the model to predict not only point estimates, but also full probability distributions for the number power outages. Thus, it provides more complete information about the uncertainty associated with the power outage predictions. These predictions then better support utility decision-making. This is the first study that predicts the probability distribution of power outages in advance of the storm. The results demonstrate the high accuracy of the QRF in effectively modeling the probability distribution of power outage data (see §2.4.1 for details).

**(3) Developing models for thunderstorm outages:** Extreme weather events include hurricanes, tornadoes, thunderstorms, snowstorms, and ice storms [58]. Thunderstorms, which occur more frequently than hurricanes, can cause power outages lasting from several hours up to several days or more. Many studies have developed predictive models for hurri-

cane power outages; however, only a few focused on thunderstorms (see Table 2.1 in §2.2.1). Although thunderstorms generally cause more frequent outages than hurricanes, their consequent outages are spatially irregular rendering them very difficult to predict. The proposed two-stage model can help utilities significantly in outage recovery for thunderstorms.

### 2.1.3 Chapter Organization

This chapter is organized as follows. Section 2.2 provides a literature review of power outage predictive modeling (POPM) and approaches proposed for learning from imbalanced or zero-inflated data. In Section 2.3, after describing the data, I introduce the probabilistic two-stage modeling framework. Section 3.4.3 provides the computational results from using the proposed method for modeling the actual data from 11 thunderstorms in Alabama. In this section, the importance and influence of different variables in modeling the power outages are also investigated. Finally, Section 2.5 concludes the chapter.

## 2.2 Literature Review

This chapter is closely related to two main domains of research, namely predictive models for storm power outages and learning from zero-inflated data.

### 2.2.1 Power Outage Prediction

**Statistical models for POPM:** A wide range of models have been developed in the literature, beginning with parametric statistical models. Han et al. [42] developed a negative binomial generalized linear model (NB-GLM) to estimate the spatial distribution of hurricane power outages. Han et al. [41] further improved the predictive accuracy of their previous work [42] by using a Poisson generalized additive model (GAM). They found that GAM can capture the nonlinearity in the data, and overcome the over-prediction problems of the NB-GLM.

Non-parametric models gained popularity shortly thereafter. Guikema et al. [38] developed non-parametric models for outage forecasting including classification and regression trees (CART), and Bayesian additive regression splines (BART). They compared the predictive accuracy of their models with the GAM and GLM, and showed that non-parametric approaches outperform the parametric ones. Later, an ensemble of tree-based models gained popularity. Kankanala et al. [58] proposed an ensemble model based on a boosting algorithm for estimating wind and lightning related power outages. They showed that boosting algorithms estimate power outages better than neural networks and a mixture of experts.

Nateghi et al. [82] developed another ensemble method based on random forests (RFs) to estimate the number of hurricane power outages. They showed that this model yields more accurate predictions than the Han et al. [42] model.

**Unbalanced data in POPM:** Xu et al. [126, 127] considered the issue of imbalanced data in power systems. They applied the *E*-Algorithm and artificial immune recognition system to identify distribution fault causes. They showed that these two algorithms offered improved performance relative to artificial neural networks when the data is imbalanced. Two-stage modeling is another approach suggested by Guikema and Quiring [37] to deal with the zero-inflation. Building on [37], McRoberts et al. [74] developed a two-stage model using RF models in the first and second stages. They showed that the two-stage approach effectively handled the zero-inflation issue and captured the complexities existing in power outage data.

**POPM for various weather events:** There exists a considerable body of research on estimating power outages caused by hurricanes (e.g., [74, 82, 83, 91, 38, 41]). But, there are only a few papers in the literature developing models for non-hurricane weather events. He et al. [44] and Wanik et al. [121] developed models for predicting outages caused by various storm events including hurricanes, blizzards, and thunderstorms. Zhou et al. [136] presented two models to estimate the failure rates of overhead power distribution lines caused by thunderstorms and ice/snow storms. Liu et al. [66] developed models based on a large data set of historical hurricane and ice-storm outages. Sarwat et al., [97] use the combined effects of common weather conditions to predict the total number of daily power distribution interruptions in a region.

**Uncertainty in POPM:** Despite more than a decade of research in storm POPM that have led to a steady reduction in forecast errors, power outage forecasts are not yet perfect and there are different types of uncertainty in the forecasts. They include uncertainty in the inputs (e.g., model structure or predictors) or outputs (e.g., estimated parameters of the models or predicted outages) of the models. Quiring et al. [90] investigated the impacts of tropical cyclone track and forecast errors on hurricane POPM using Monte Carlo simulation. They show that small errors in the official track and/or intensity forecast lead to large errors in the resulting outage predictions. He et al. [44] developed two models based on BART and quantile regression forest (QRF) for obtaining prediction intervals. They find that the BART model predicts more accurate point estimates, but the QRF makes better prediction intervals. Table 2.1 summarizes the power outage predictive modeling literature.

Table 2.1: Literature review table for storm power outage predictive modeling

Paper	Model	Probabilistic/ deterministic			Storm type					unbalanced	Response variable	spatial unit (km*km)	Region of study
		Uncertain input	Uncertain output	Deterministic	Hurricane	thunderstorm	Ice storm	Blizzards	Not specified				
[136]	Poisson-regression; Bayesian network	✓							✓		Power outage	Circle of 9 mile radius	Manha- ttan,KS
[127]	AIRS			✓					✓	✓	Outage causes	Each outage	NC, SC
[126]	E-Algorithm			✓					✓	✓	Outage causes	Each outage	NC, SC
[66]	AFT			✓	✓			✓			Power restoration time	Zip code	NC, SC, VA
[94]	Combined statistical-GIS method			✓	✓	✓					Outage duration fragilities and restoration	Entire area affected by storms	Seattle
[67]	GLMM			✓	✓			✓			Power outages	Zip code & 3*3	NC, SC, VA
[41]	GLM(NB)			✓	✓						Power outages	3.66*2.44	GCR
[42]	GAM			✓							Power outages	3.66*2.44	GCR
[38]	GLM(NB); GAM CART; BART			✓	✓						Damaged poles	3.66*2.44	Missi- ssippi
[91]	CART			✓	✓						Power outages	3.66*2.44	Central GCR
[83]	CART; BART; MARS			✓	✓						Outage duration	3.66*2.44	Central GCR
[37]	CART-GAM			✓	✓					✓	Power outages	3.66*2.44	GCR
[58]	AdaBoost			✓					✓		Power outages	Entire city	KS
[90]	CART-GAM	✓			✓					✓	Power outages	3.66*2.44	Central GCR
[82]	RF			✓	✓						Power outages	3.66*2.44	Central GCR
[81]	RF			✓	✓						Outage duration	3.66*2.44	Central GCR
[39]	RF			✓	✓						Power outages	3.66*2.44	U.S. Coastline
[121]	DT; RF; BT; DT+RF			✓	✓	✓		✓			Power outages	2*2 & 6*6 & 18*18	Conne- cticut
[74]	RF-RF			✓	✓					✓	Power outages	3.66*2.44	U.S. Coastline
[44]	BART; QRF		✓	✓	✓	✓		✓			Power outages	2*2	Conne- cticut
My model	RS-RF-QRF RF-QRF CS-RF-QRF BT-QRF RS-BT-QRF SVDD-QRF CS-BT-QRF SVM-QRF RS-SVM-QRF		✓	✓		✓				✓	Customer outages	3.66*2.44	AL

AFT: accelerated failure time; GLMM: generalized linear mixed model; MARS: multi additive regression splines  
 BT: boosting trees; DT: decision trees; RS: resampling; CS: cost-sensitive; SVM: support vector machine  
 SVDD: support vector data description; GCR: Gulf Coast region; NC: North Carolina; SC: South Carolina  
 KS: Kansas; VA: Virginia; AL: Alabama

### 2.2.2 Learning from Zero-inflated or Imbalanced Data

Zero-inflated models are mostly based on a two-stage process and are divided into *hurdle* and *mixture* models.

**Hurdle models:** In the first stage of a hurdle model, a binomial probability formulation models whether the outcome variable has a zero or non-zero value. If the first stage model determines that the realization is non-zero, then the conditional distribution of the non-zero realizations is modeled with a zero-truncated model. Building upon the hurdle model, zero-altered Poisson is built in which zero observations are modeled with a binomial distribution and the non-zero observations are modeled with a truncated Poisson model.

**Mixture models:** In the first stage of a mixture model, instead of modeling all zeros, only a proportion of them are modeled with a classifier. The other part of the data which is not yet labeled as zeros, are considered as the second population. Then another model is fit to the second population. This model produces zeros as well as non-zeros. In zero-inflated Poisson (ZIP) [63], a binomial-GLM fits to the data to model the probability of being zero, then a Poisson-GLM is used to model the count process. The main difference between the mixture and hurdle models is that unlike the hurdle models, the count process produces zeros in the mixture models, which results in more flexibility.

Since ZIP and zero-inflated negative binomial (ZINB) cannot always explain the performance of the system adequately, Guikema and Quiring [37] proposed a mixture model for predicting zero-inflated power outages. The first stage is a CART predicting whether the outage is zero or not, and a Poisson-GAM model is used for the second stage to predict the number of outages. This model improves the accuracy of predictions over the ZIP and ZINB models. Building on [37], McRoberts et al. [74] proposed the use of a random forest model for the first and second stages of their mixture model and improved the accuracy of the model significantly.

**Classifying unbalanced data:** In above-mentioned models, there is a binary classifier in the first stage, which explains whether the response variable is zero or not. Having an accurate classifier helps improve the overall performance of the two-stage model. Standard classification algorithms assume that the number of observations from different classes is roughly similar while in zero-inflated data, a high proportion of records is zeros. To overcome the challenges arising from imbalanced data, two main approaches are used including (1) algorithm-level methods, and (2) data-level methods [61].

**Algorithm-level approaches:** Algorithm-level approaches concentrate on modifying the existing ML models to reduce their bias towards the majority class and improve their performance. A common approach is cost-sensitive methods in which higher costs are assigned to the prediction error of the minority class [22]. This approach boosts the importance

of the minority class and alleviates the bias towards the majority class. Since their computational efficiency is higher than resampling methods, they are more suitable for large data sets [40]. Another approach is one-class learning which is used to capture the properties of the minority class. To train a one-class learning model, two strategies are followed. In the first, only the objects from the target class are used to train a model describing the target set and objects from another class are ignored. In the second, examples of both classes are used, although the focus is more on accurate predictions for the minority class [76].

Ensemble methods combine several base learners to improve the performance of any single one. They have become a popular method for learning from imbalanced data. These methods are categorized into iterative based ensembles (e.g., boosting methods), and parallel based ensembles (e.g., bagging and RF). Galar et al. [31] present a survey of using ensemble methods for imbalanced learning. The performance of an ensemble model is affected by the accuracy of the base learners, and diversity between all the learners [118].

**Data-level approaches:** Data-level approaches include resampling and feature selection. Resampling is a method for rebalancing the training set to reduce the effect of the majority class. They are independent of the selected classifier [40] and fall into three groups: under-sampling, over-sampling, and hybrid methods. In under-sampling, it is assumed that many instances from the majority class are redundant and so, some are discarded to make the training set roughly balanced. In over-sampling, new examples from the minority class are created. Hybrid methods are a combination of over-sampling and under-sampling. Resampling can be done randomly, or based on some strategies (e.g., clustering-based, distance-based, and evolutionary-based). Feature selection is, in general, selecting a subset of variables among all potential predictive variables to allow a learning algorithm to achieve optimal performance. It has been used less than resampling methods for imbalanced data [40]. It can improve the predictive accuracy and reduce the bias toward the majority class because the irrelevant features might cause the model to discard the minority class examples as noise [132].

## 2.3 Data Description and Methods

### 2.3.1 Data Description

I use data from 11 strong thunderstorms that have occurred over the past ten years in the state of Alabama. This area is divided into 6,623 3.66 km by 2.44 km grid cells (see Figure 2.1-a). The variables are divided into two categories. The first is related to the power system, geographic characteristics, and tree and soil characteristics which are time-invariant. The second contains the variables that are time dependent and represents the

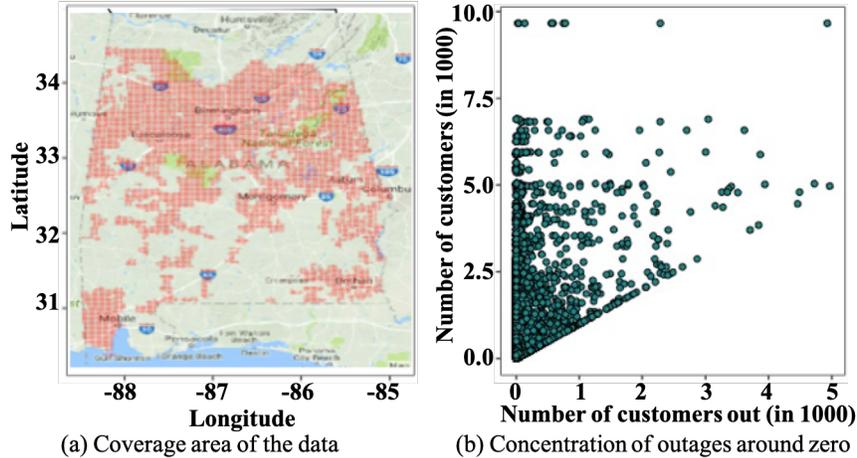


Figure 2.1: The red area in the left figure shows the total coverage area of my data set. The right figure plots the number of customers vs number of customers with power outages for all the instances. It shows the concentration of customer outages around zero, and zero-inflation issue in the data.

pre-storm conditions such as various measures of soil moisture, precipitations and weather forecasts. The data originate from different sources and have different spatial resolutions. All data are converted to the level of grid cells, and the number of customers without power are predicted for each cell. Each of these variables is explained below.

**Geographic variables:** Geographic variables include measures of topography and land cover. The topographical variables are collected from a global 30-arcsec digital elevation model produced by the U.S. Geological Survey. They include the min, max, mean, standard deviation, and median of elevation. The land cover (LC) variables are from the National Land Cover Database. They are summarized into eight major classes including water, developed, barren, forest, scrub, grassland, pasture, and wetlands.

**Tree and soil characteristics:** The eight tree-related variables are collected from 2012 National Insect and Disease Risk Map. They include fractional area of a grid cell covered by trees, percentage of deep-rooted trees, percentage of tap-rooted trees, maximum diameter at breast height, maximum height, wood density, Janka Hardness Scale, and crushing strength of trees. Root zone depths variables are derived from the USDA Gridded Soil Survey Geographic. They are defined as the depth within the soil column from which roots can extract water [74].

**Power system variables:** To characterize the power system, I include the number of poles, switches, transformers, and total length of overhead and underground line in each grid cell. They provide a measure of the extent of power system exposure to high winds.

**Soil moisture and Precipitation:** These variables, derived from the North America

Land Data Assimilation System, are measures of local drought and soil moisture prior to the storm. They help to explain the stability of poles and trees. Soil moisture is estimated at three depths: 0-10 cm, 10-40 cm, and 40-100 cm. The standardized precipitation index (SPI) is a measure of precipitation deviations from normal conditions. SPI is estimated for different duration: 1, 3, 6, 12, and 24 months (e.g., a 3-month SPI is a measure of the deviation of precipitation from the long-term average in the 3 months prior to a storm).

**Weather forecasts:** The weather data are obtained from the National Digital Forecast Database and include dewpoint, temperature, relative humidity, sky cover, air temperature, 2-minute wind speed, maximum instantaneous wind gust (m/sec) and wind direction. The data also include probability of a tornado within 25 miles, 12 hour(h) probability of precipitation, 24 h quantitative precipitation forecast, probability of winds greater than 58 mph within 25 miles, probability of Enhanced Fujita scale 2 tornadoes within 25 miles, risk of fire, daily probability of a convective hazard, probability of hail greater than 0.75 and greater than 2 inches in diameter within 25 miles, and probability of winds greater than 75 mph within 25 miles. I also included binary variables that indicate whether the NWS has issued a flash flood, severe thunderstorm, and tornado watch for a given grid cell.

**Response variable:** The response variable for my model is the number of customers without power in each grid cell. This data comes from a combination of customer call-in data and a model of the electric power system that estimates which customers would be without power given the activated protective devices and customer call-ins. The number of customer outages is highly zero-inflated. Over 90% of grid cells have no power outages in my data set. Figure 2.1-b illustrates the number of customer outages versus the number of customers by grid cell. The density of instances equal to zero indicates the zero-inflation property of the data.

**Variable selection:** Variable selection is an important task especially when the number of variables is large and the data set is unbalanced [61]. Since the original data set contains many covariates that are highly correlated with each other, variable selection can help obtain a simpler model and improve accuracy. My approach for variable selection is to find a subset of covariates in which there is no significant sign of collinearity. Collinearity does not necessarily harm the predictive accuracy; however, variables which have high collinearity with others can potentially be removed without deteriorating model performance. Further, it can simplify the model and make the interpretation easier. In this study, the correlated covariates are found using the pairwise correlation plots, and calculating the variance inflation factor (VIF) of each covariate. The covariates with high VIF can be explained by linear combinations of other covariates. By considering both correlation plots and VIF values, I then remove the covariate with the highest collinearity with others in an iterative process.

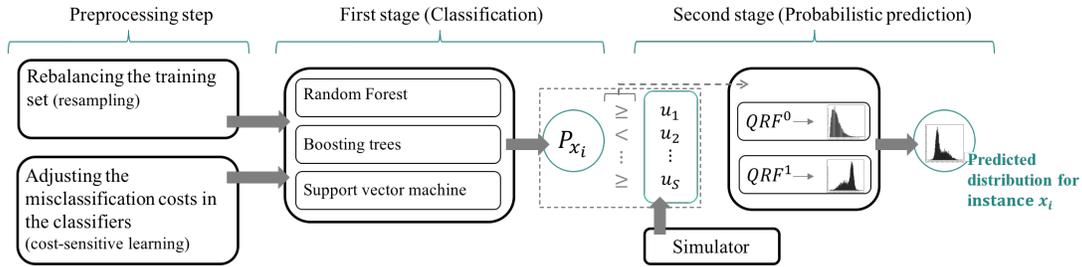


Figure 2.2: The illustration of the proposed two-stage framework. In the preprocessing step, either training data are rebalanced or misclassification costs of the non-zero class instances are enhanced in the classifier. In the first stage, a classifier predicts the probability of each instance,  $x_i$ , having non-zero outages,  $P_{x_i}$ . In the second stage, firstly, a simulator generates random numbers between 0 and 1,  $u_i, \forall i = 1, \dots, S$ ; then,  $QRF^0$  and  $QRF^1$  predict two separate distributions for each instance by assuming that the instance belongs to the zero or non-zero class respectively. Finally, by considering the values of  $P_{x_i}$  and  $u_i$ s, random records are generated from the predicted distributions. These records together estimate the full probability distribution for each instance  $x_i$ .

This results in a set of 32 covariates. All of the models described in the following section are trained on this reduced set of variables.

### 2.3.2 Models Implemented

In this study, a novel two-stage framework (see Figure 2.2 for graphical presentation) is proposed to model the distribution of power outages prior to a storm. In this framework, a classifier is embedded in the first-stage predicting the probability of having at least one customer without power in each spatial unit. In this regard, three state-of-the-art classifiers (RF, BT and SVM) are chosen. Each of these methods is explained later in this section. Aiming to overcome the zero-inflation issue and boost the accuracy of these classifiers toward non-zero class data, resampling and cost-sensitive learning are used. In the former case, the classifier is trained on the rebalanced data; while in the latter one, different misclassification costs are selected for the record belonging to the zero and non-zero classes. Applying cost-sensitive learning is different for each of the RF, BT and SVM classifiers and later it is explained for each classifier. In the second stage of my framework, I developed two QRF models, one for those grid cells classified as zero and one for those grid cells classified as non-zero. Using these models, and the predicted probability from the first stage, the full probability distribution for the number of customer outages in each grid cell is predicted.

I have developed 9 two-stage models in total. The second stage of all these models is QRF. However, the first stage and pre-processing step of each model are different. It is a combination of one of the pre-processing steps (resampling, cost-sensitive learning or none of

them) and RF, BT or SVM. So, I shall call these 9 two-stage models RF-QRF, Resampled-RF-QRF, Cost-sensitive-RF-QRF, BT-QRF, Resampled-BT-QRF, Cost-sensitive-BT-QRF, SVM-QRF, Resampled-SVM-QRF, and SVDD-QRF. In the following, each component of my two-stage framework is explained in more detail.

### 2.3.2.1 Resampling

To rebalance the data, I apply two techniques (1) random under-sampling in which many instances from the zero class data are removed randomly, and (2) Synthetic Minority Over-Sampling Technique (SMOTE). SMOTE [11] is a popular approach in which, rather than replicating minority class records, synthetic instances are generated and added to the original data. It can help to avoid over-fitting. Both random under-sampling and SMOTE are successful techniques commonly applied to the class imbalanced problems [6]. In section 3.4.3, their effectiveness in improving the accuracy of outage prediction is evaluated.

### 2.3.2.2 Random Forest

RF [8] is an ensemble model in which many trees are trained on the bootstrapped data and the output is the average of trees' predictions. At each node of a classification tree, the best splitting variable and point are picked from a set of variables selected randomly from all variables aiming to reduce the impurity in each node. RF uses both bagging and random variable selection for tree building, which results in low correlation between individual trees. As a result of low bias and variance, RF often yields strong predictive accuracy. For regression, RF uses the sum of squared errors as the impurity measure. Moreover, for each terminal node of a grown tree, only the mean of the response values is kept and all other information of the instances are neglected. This mean value is represented as a prediction for any instance belonging to the corresponding leaf. In contrast, QRF [75], which is a generalization of RF model, keeps the value of all observations in each leaf. QRFs therefore consider the spread of the response variable and estimate any quantile of the response variable. QRF is selected for the second stage of my models because it is appropriate for producing probabilistic predictions, and the initial analysis shows its high performance in effectively modeling the power outage data.

I develop Resampled-RF-QRF and Cost-sensitive-RF-QRF models. The first trains the RF on the rebalanced data using resampling. The second is based on a cost-sensitive RF in which higher cost is assigned to misclassification of the minority class (non-zero outages) by using a weighted Gini index as node impurity function, and minimize the overall cost. Performance of these models is compared together and with RF-QRF and results are shown

in §3.4.3.

### 2.3.2.3 Boosting Trees

Boosting [98] combines the performance of many weak learners to improve predictive power. AdaBoost [30] is a commonly used boosting method. AdaBoost calls a weak learner, which can be any statistical model, in a series of rounds and in each round it provides the weak learner with the distribution  $D_t$ , which is updated in each round for any instance. Initially, the same weight is given to all instances. In next rounds, for any correctly classified instance,  $D_t(i)$  is decreased; however, it is increased for the incorrectly classified ones. Thus, the easy instances that are classified correctly in many of the rounds get lower weight, while the hard ones that are mostly misclassified get higher weights [30].

There are multiple ways to introduce the cost items to the AdaBoost, and the most common ones are AdaC family [108], AdaCost [25], and CSB [111]. Because a preliminary study indicates that the AdaC2 outperforms other AdaC methods for separating zero outages from non-zeros, I use the AdaC2 method as the cost-sensitive BT classifier. Thus, I develop Cost-sensitive-BT-QRF mixture model in which the first stage is a AdaC2 and the second stage is a QRF.

### 2.3.2.4 Support Vector Machine

SVM [17] combined with kernel techniques can be used for classification of both linear and non-linear data. It maps the original data to a higher dimension, where a maximal distance hyper-plane can be found as a discriminant function for the separation of data using instances called support vectors. The determination of an optimal hyperplane leads to solving the following optimization problem by using Lagrangian multiplier  $\alpha_i$ :

$$\begin{aligned} \min_{\alpha_i, \alpha_j \geq 0} \quad & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^n \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0, \quad \forall i = 1, \dots, n \end{aligned}$$

where  $x_i$  and  $y_i$  are predictors vector and response variable respectively, and  $k(x_i, x_j)$  is the kernel matrix. The choice of kernel function affects the model performance, and the common kernels are linear, polynomial, Gaussian and sigmoid. I use the Gaussian kernel because it is by far one of the most powerful ones and develop Resampled-SVM-QRF and SVM-QRF models in which the first is trained on the rebalanced data, and the second is trained on the original data.

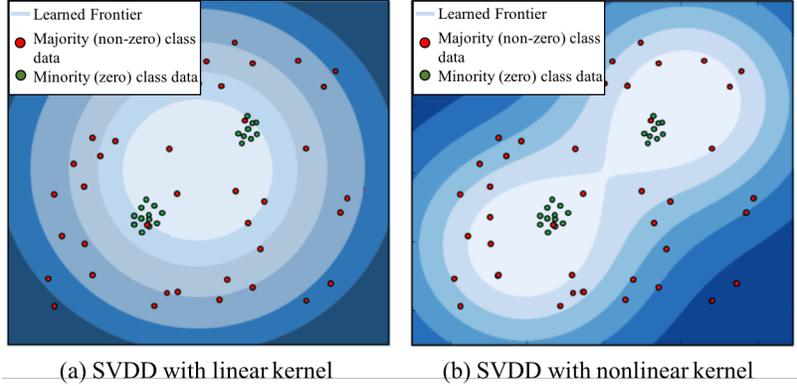


Figure 2.3: Comparison of the boundaries obtained by SVDDs with linear and nonlinear kernels.

A SVM classifier is also used in the context of one-class learning for unbalanced data, which is called Support Vector Data Description (SVDD) [110]. Instead of partitioning the space with boundaries, SVDD makes a hypersphere which surrounds the target class (usually the minority class). The hypersphere is learned by using data of both classes or data of the single target class. Using non-linear kernels in the SVDD results in more flexible and tighter hypersphere around the target class. Figure 2.3 indicates the boundaries obtained by SVDDs with linear and nonlinear kernels. It shows that SVDD with nonlinear kernel is more flexible to build a separating boundary around the target class. In this study, I employ the SVDD method in the first stage of my mixture model as a classification technique. My analysis indicates that the SVDD with Gaussian kernel yields more accurate boundaries. This method finds optimal boundaries around non-zero class data and predicts whether each instance belongs to the boundary (i.e., non-zero), or not (i.e., zero). Second stage of this model is a QRF and I call it SVDD-QRF.

## 2.4 Computational results and analysis

Using the proposed framework, nine mixture models are developed to predict distributions and point estimates of thunderstorm power outages. The number of trees in QRF is 500 and the minimal number of instances in each terminal node is 50 to generate a large number of quantiles. The models produce 101 percentiles (0%, 1%, ..., 100%) as the predicted distribution for each grid cell and the means of these predicted distributions are calculated as point estimates. To compare the models, 10 random holdout validation tests are conducted. In each, I randomly hold out 20% of the data as the validation set, leaving the remaining as the training set. The model is developed on the training set, and tested on the validation

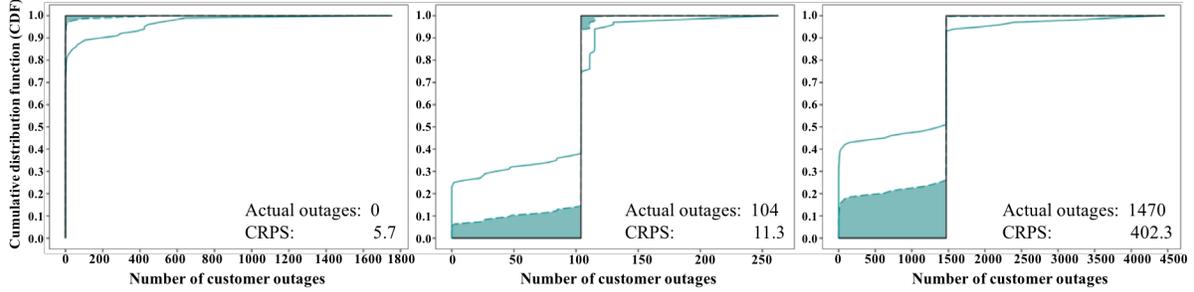


Figure 2.4: Continuous Ranked Probability Score (CRPS) for three typical grid cells with observed outages equal to (a) 0, (b) 104, and (c) 1470 for the storm occurred in April 2006. The green solid curve shows the predicted CDF, and the black vertical line shows the actual value for the number of customers without power for each grid cell examples. The shaded green area under the green dashed curve is the CRPS for the predicted CDF. CRPS is a measure for presenting the variation of the predicted CDF from the actual value. The closer the green solid curve is to the black stepwise function, the more accurate is the predicted CDF and smaller is the CRPS value.

set. In each iteration, the performance metrics are calculated for each model.

In Resampled-RF-QRF, Resampled-BT-QRF, and Resampled-SVM-QRF models, the training set is rebalanced by randomly removing half of the records with zero outages (random undersampling), and doubling the number of records with non-zero outages by adding new non-zero records using SMOTE. In Cost-sensitive-RF-QRF model, the cost of misclassifying the records with outages larger than 100 is set four times of the cost of the other records. For Cost-sensitive-BT-QRF, the misclassifying cost is set to 4 for the records with outages larger than 100, and 1 for the other records. In SVDD-QRF, the SVDD is trained by using the data of only non-zero outages. The nine mixture models are also compared with the null model, in which the predicted distribution of outages in the validation set is set to the distribution of outages in the training set. Note that all model parameters are chosen based on cross-validation (C.V.) technique to maximize the accuracy of these nine mixture models.

#### 2.4.1 Probabilistic and Point Predictions

**Training phase:** In all nine models, the first stage model is trained on the training set where the response is binary indicating whether the outage is zero or not. I call this model  $f^1(x)$ . Using this model, prediction for the training set is made and the output is the probability of each record being non-zero. By specifying a threshold assumed to be 0.6 (set through C.V.), the training set is divided into two sets, zeros and non-zeros. For any  $x_i$ , if  $f^1(x_i) > 0.6$  then  $x_i$  is added to the non-zero class and the zero class otherwise. Then two separate QRF models are trained on these sets which are called QRF<sup>1</sup> (the non-zeros set)

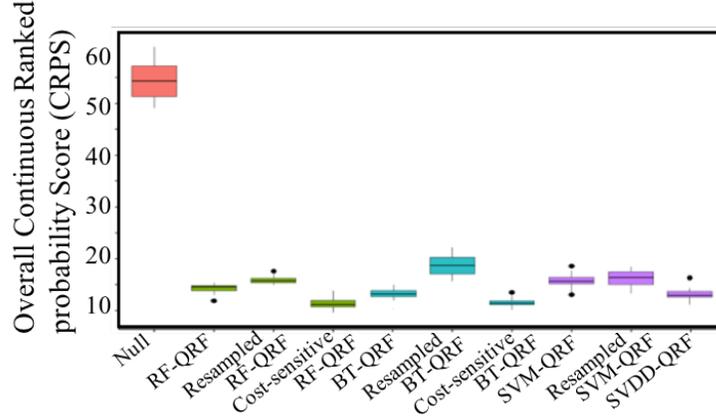


Figure 2.5: Comparing probabilistic predictions of all models based on CRPS

and  $QRF^0$  (the zeros set).

**Prediction phase:** To make a probabilistic prediction for a grid cell in the validation set, I follow the procedure in Algorithm 1 where  $S$  is set to 10,000. This approach helps us not assign the records for which the output of the first stage model is close to the threshold deterministically to one class.

**Continuous Ranked Probability Score (CRPS):** To evaluate the probabilistic mixture models, CRPS [34] is used. CRPS generalizes MAE to the case of probabilistic forecasts. It is defined as

$$CRPS(F, y_0) = \int_{-\infty}^{\infty} (F(y) - \mathbb{1}\{y \geq y_0\})^2 dy$$

where  $F$  is the predicted cumulative density function (CDF) for the response value  $y_0$ . In hold-out analysis, the CRPS is calculated for each record in the validation set, and the mean

---

**Algorithm 1** Probabilistic Prediction by Two-stage Modeling Framework

---

- 1: Initialize the classifier,  $QRF^1, QRF^0, S$  and the validation set.
  - 2: **for** each instance  $x_i, i = 1, \dots, T$  in the validation set, **do**
  - 3:     Predict probability of non-zero response using the classifier,  $P_{x_i}$ .
  - 4:     Using  $QRF^0$  predict a CDF for the response variable,  $CDF_0$ .
  - 5:     Using  $QRF^1$  predict a CDF for the response variable,  $CDF_1$ .
  - 6:     **for**  $j = 1, \dots, S$  **do**
  - 7:         Generate a random value,  $u_j \sim \text{Unif}(0, 1)$ .
  - 8:         **if**  $P_{x_i} > u_j$  **then**
  - 9:             Generate a random record from  $CDF_1$ .
  - 10:         **else:**
  - 11:             Generate a random record from  $CDF_0$ .
  - 12:     Estimate CDF of response variable by all generated records together.
  - 13:     Output the mean of generated records as the point estimate for the response variable.
-

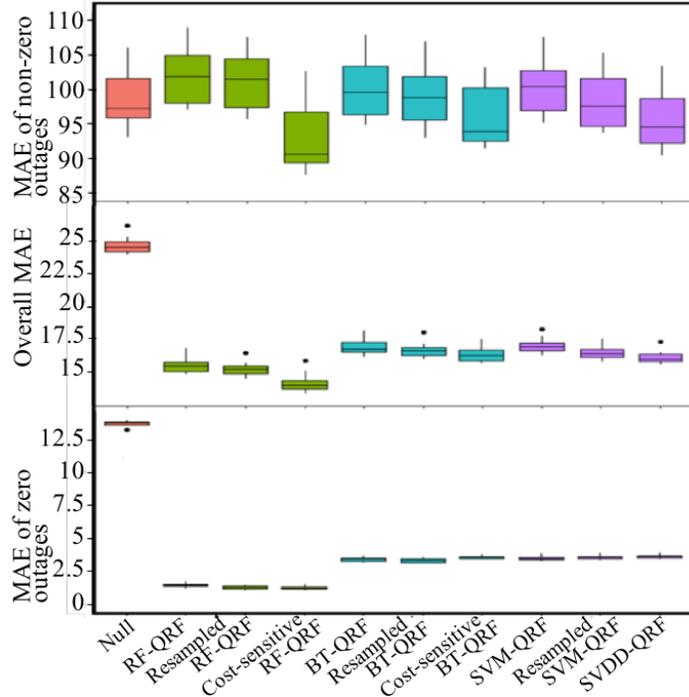


Figure 2.6: Comparison of the point estimates obtained by different models in terms of the MAE. The overall MAEs as well as the MAEs of the records with zero outages and records with non-zero outages are calculated separately.

of all CRPSs over the validation set is considered as the mean CRPS for the corresponding model and iteration. Figure 2.4 illustrates the predicted CDF and the actual number of outages for three typical grid cells. In this figure, the green shaded area presents the CRPS for each predicted CDF. It also shows how CRPS can capture the variation of the predicted distributions from its actual value.

Figure 2.5 compares the models by their CRPS in the hold-out analysis. It shows that all mixture models predict distributions of outages significantly better than the null model. Applying cost-sensitive learning in Cost-sensitive RF-QRF has improved the probabilistic predictions significantly. However, applying the resampling technique has deteriorated the performance in almost all three types of models. This may be because resampling changes the distribution of the training set, and so the predicted distributions do not match the data very well.

The mean of predicted distributions is calculated as a model’s point predictions. To evaluate the model’s point predictions, mean absolute error (MAE) is estimated. Since my data is highly zero-inflated, the overall MAE as well as the MAEs of records with zero outages and records with non-zero outages are compared separately. Figure 2.6 summarizes the accuracy of all models. Comparing the MAE for non-zero records shows that both resampling and cost-sensitive methods are effective in improving the model performance for predicting non-

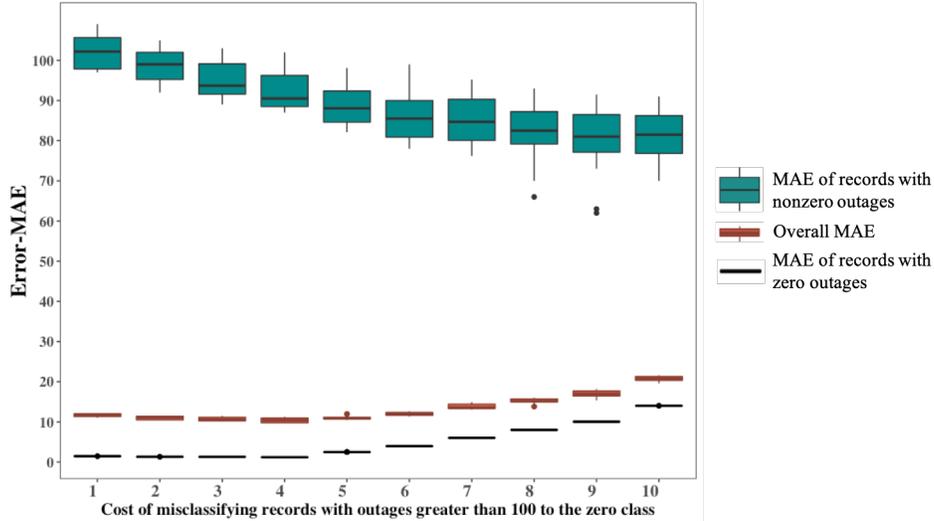


Figure 2.7: Analyzing the accuracy of cost-sensitive RF-QRF model for difference values of the cost parameter.

zero outages while they do not deteriorate the model performance for predicting zero outages. They handle the zero-inflation issue better than the two-stage modeling approach without the preprocessing step proposed by [74] and predict power outages with higher accuracy. Figure 2.6 shows that the mixture models with a RF in the first stage perform significantly better than others for predicting both zero and non-zero outages. It also illustrates that all 9 models estimate outages with significantly greater accuracy than the null model. Table 2.2 summarizes the MAE and CRPS values for all models.

Since the cost-sensitive RF-QRF performs better than other models in predicting the

Table 2.2: Comparing the accuracy of various models

Model	Probabilistic Predictions		Point Predictions	
	CRPS	MAE Zero class	MAE Nonzero class	MAE Overall
RF-QRF	14.4	1.4	102	15.4
Resampled RF-QRF	15.8	1.3	101.1	15.2
Cost-sensitive RF-QRF	<b>11.1</b>	<b>1.2</b>	<b>92.9</b>	<b>14.1</b>
BT-QRF	12.0	3.4	100.1	16.9
Resampled BT-QRF	17.6	3.3	98.9	16.6
Cost-sensitive BT-QRF	11.5	3.5	96.5	16.4
SVM-QRF	15.4	3.5	100.3	17.8
Resampled SVM-QRF	16.3	3.6	98.4	16.5
SVDD-QRF	14.5	3.6	96.6	16.5
Null	54.5	13.7	100	24.6

outages, I next evaluate the impact of the misclassification cost in its first-stage classifier on the model’s performance. Figure 2.7 indicates the overall MAE as well as the MAEs of the records with zero outages and records with non-zero outages for 10 different misclassification costs. These values are the costs of misclassifying the records with outages greater than 100 to the class of records with zero outages, while the misclassification cost for all other records is set to 1. As the cost of misclassifying the records with outages greater than 100 increases, the MAE for the non-zero class records decreases. On the other hand, the overall MAE and MAE for the zero-class records do not change significantly for small cost values. However, these errors increase as the cost parameter increases. Choosing the cost parameter depends on the policies taken by the decision maker and the trade-off between these three types of errors. Large error for the zero-class records results in over-estimating the outages and causes higher costs of preparation for utility companies, while under-estimating the outages results in higher customers dissatisfaction. In this analysis, the main purpose is to have the minimum overall MAE and thus, I choose the cost of 4 (which has the lowest overall MAE) for the records with outages greater than 100.

## 2.4.2 Partial Dependence Plot and Variable Importance

A partial dependence plot (PDP) illustrates the marginal contribution of a single variable to the response. I use PDPs (Algorithm 2), presented by [51], for assessing the influence of variables in the two-stage model. In this algorithm,  $x_i = (x_{i1}, x_{i2}, \dots, x_{im}) \quad \forall i = 1, \dots, n$  denotes  $i^{th}$  instance from the data set. Given the lowest CRPS and MAE in the hold-out analysis, the cost-sensitive-RF-QRF model performs the best, and so it is trained on the whole data and used as a reference model for making PDPs. Based on this model, I plot the PDPs. Figure 2.8 shows PDPs for the 9 most influential variables. A relatively flat PDP

---

### Algorithm 2 Partial Dependence Plots and Variable Importance

---

- 1: Initialize the predictive model  $f(x_i)$  trained on the whole data, number of covariates  $m$ , and total number of instances  $n$  in the whole data.
  - 2: **for** each variable  $x_j, j = 1, \dots, m$  **do**
  - 3:     Define a range with discrete values for the covariate to iterate over     i.e.,  $(a_{1j}, a_{2,j}, \dots, a_{Kj})$ .
  - 4:     **for** each  $a_{kj}, k = 1, \dots, K$  **do**
  - 5:         **for**  $i = 1, \dots, n$  **do**
  - 6:             Substitute  $a_{kj}$  with  $x_{ij}$  within  $x_i = (x_{i1}, \dots, x_{ij}, \dots, x_{im})$      (new instance is denoted by  $\bar{x}_i$ ).
  - 7:             Make prediction for  $\bar{x}_i$  (output is  $f(\bar{x}_i)$ ).
  - 8:             Calculate  $f_{a_{kj}} = \frac{1}{n} \sum_{i=1}^n f(\bar{x}_i)$ .
  - 9:             Plot PDP for variable  $k$  (i.e., plot  $f_{a_{kj}}$  against  $a_{kj}$ ).
  - 10:            Find  $M_j = \max_k(f_{a_{kj}})$ , and  $m_j = \min_k(f_{a_{kj}})$ .
  - 11: Estimate the importance of each variable as  $VI_j = \frac{M_j - m_j}{\sum_{j=1}^m M_j - m_j}$ .
-

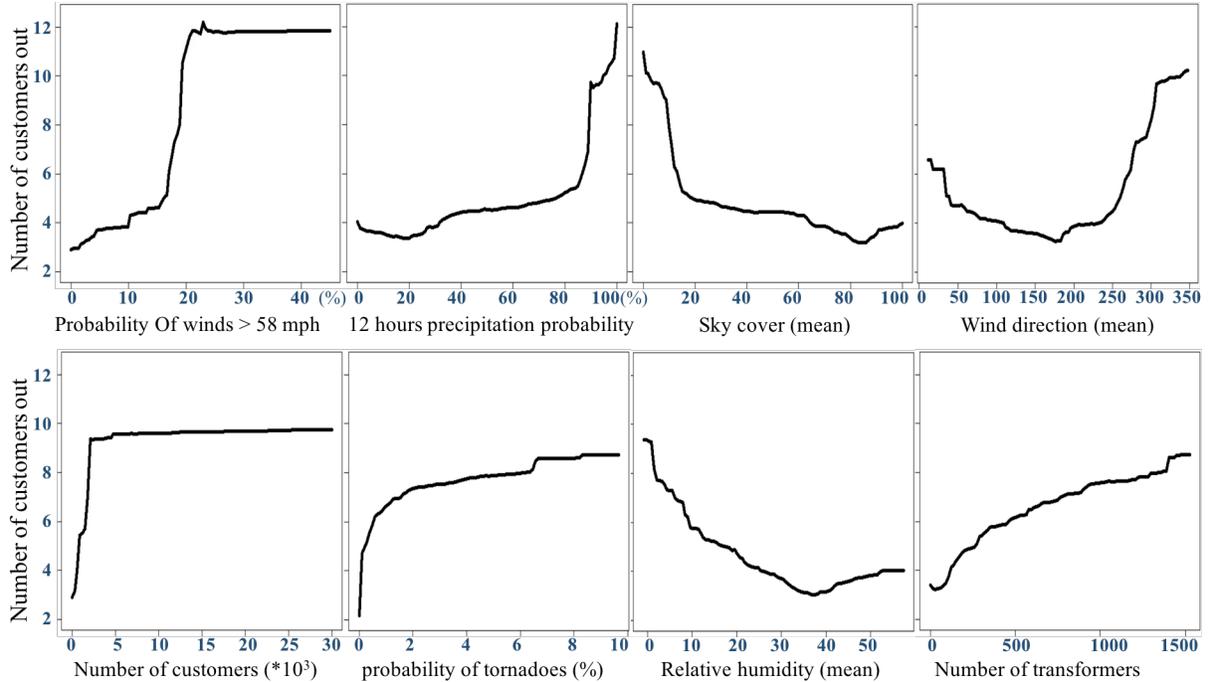


Figure 2.8: Partial dependence plot for the 8 most important variables

indicates that the covariate of interest has relatively little influence on the predictions, while a large variation indicates that the covariate has a large degree of effect on the predictions. Figure 2.8 indicates that the number of outages has a direct relationship with probability of winds being greater than 58 mph, 12 hour precipitation probability, probability of tornadoes, number of customers and number of transformers. Increases in these variables cause the mean of outages to also increase. However, sky cover and mean value of relative humidity have reverse relationship with the number of outages. Also, PDPs for the wind direction indicate that as the wind direction changes from southerly (wind direction of 180 degrees) to northerly (wind direction of 0 or 360 degrees), the number of outages increases steadily.

Variable importance (VI), on the other side, is a single metric based on which variables can be ranked according to their importance and influence. Algorithm 2 explains how VI is calculated and the relationship between PDP and VI. Table 2.3 shows the VI for each covariate based on its relative swing in partial dependence values. The VI values are normalized by giving the most important variable a score of 100. High VI values implies that the covariate has a large influence on predictions. As observed from Table 2.3, the weather data, number of customers and power system components are the most important types of variables. However, other types of variables like land cover, soil moisture and characteristics, tree related variables, and precipitation are less important.

### 2.4.3 Prediction for New Storms

In order to visualize the model’s predictions, the distribution of outages for one thunderstorm occurred in April, 2006 is predicted using the final cost-sensitive-RF-QRF model. Figure 2.9 shows the spatial variation of actual outages (Figure 2.9-a) versus the mean estimated outages (Figure 2.9-b), and the width of 90% confidence intervals for the estimated outages (Figure 2.9-c). Figure 2.9 shows that the overall predicted outages follow the same

Table 2.3: Variable importance in the final cost-sensitive-RF-QRF model

Rank	Type	Variable	VI
1	Weather data	Prob. of winds > 58 mph	100.00
2	Weather data	12 hours Prob. of precipitation	93.97
3	Weather data	Sky cover (mean)	83.77
4	Weather data	Wind direction (mean)	74.85
5	Weather data	Prob. of tornadoes	73.52
6	Customers	Num. of customers	70.62
7	Weather data	Relative humidity (mean)	68.07
8	Power system	Num. of transformers	59.22
9	Weather data	Wind direction (min)	46.34
10	Weather data	Precipitation forecast	42.31
11	Weather data	Convective hazard outlook	34.59
12	Weather data	Wind direction (max)	32.99
13	Power system	Total mileage of overhead lines	30.77
14	Weather data	Apparent temperature(mean)	30.27
15	Power system	Num. of poles	29.37
16	Weather data	Wind gust speed (mean)	27.86
17	Precipitation	12 month SPI	25.28
18	Power system	Num. of switches	24.94
19	Weather data	Severe thunderstorm watch	23.12
20	Topography	Elevation (mean)	20.23
21	Land cover	Forest land cover	17.58
22	Soil moisture	Soil moisture (10-40cm depth)	17.21
23	Tree	Area covered by trees (%)	17.15
24	Precipitation	6 month SPI	15.59
25	Land cover	Grassland land cover	13.56
26	Land cover	Water land cover	12.94
27	Tree	Janka hardness of tree species	11.97
28	Soil moisture	Soil moisture (0-10cm depth)	11.73
29	Land cover	Scrub land cover	11.14
30	Soil characteristic	Root zone (mean)	10.71
31	Land cover	Barren land cover	10.23
32	Tree	Trees height (max)	6.95

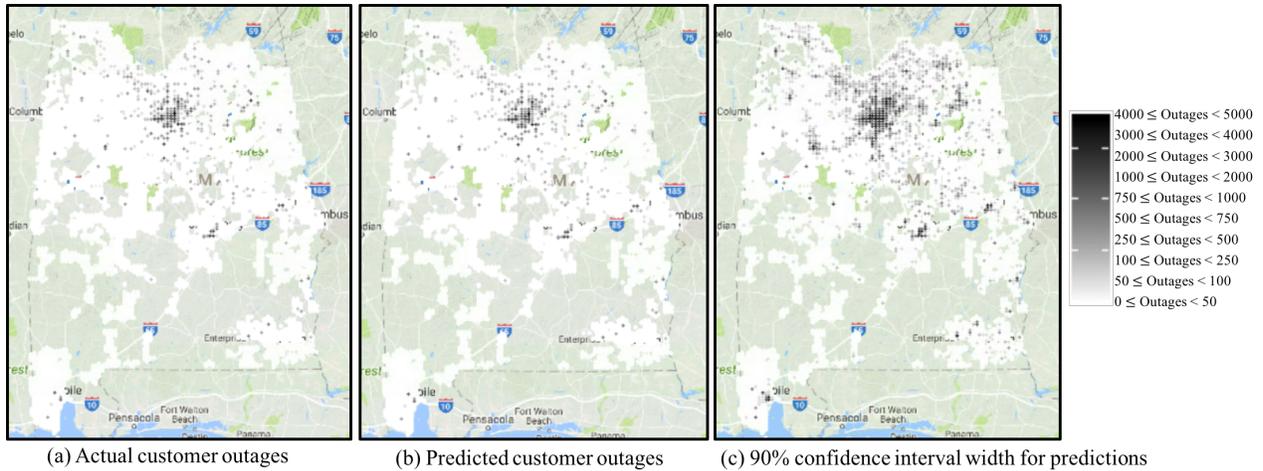


Figure 2.9: Comparison between spatial distribution of actual outages and estimated ones for the storm occurred in April 2006

general spatial pattern as the observed outages. Based on Figure 2.9-c, the estimated 90% confidence intervals for the number of outages are wider in the central and northern parts of the state relative to the other portions of the state. This might happen because these areas, which are generally associated with urban areas where a large number of customers are living, have higher concentrations of power outages.

## 2.5 Summary and Conclusions

In this study, a novel two-stage model that accurately predicts the distribution of thunderstorm-induced power outages is developed. The first stage is based on RF, BT and SVM classifiers. To deal with the zero-inflation and also improve the predictive performance of these classifiers, two techniques including cost-sensitive learning and resampling are compared.

In the second stage, there are two QRF models one of them trained on the zero class data and another one trained on the non-zero class data. Conditioning on the fact that each record belongs to the zero or non-zero class data, each QRF makes a separate prediction for the full distribution of that record. The role of the first-stage classifier is to predict the probability of the outcome variable being non-zero. Once this probability is estimated, a large number of random samples between 0 and 1 are generated. Then each random sample is compared with the probability of the outcome being non-zero. For each random sample larger than the estimated probability, a data point is randomly generated from the predicted distribution by the  $QRF^0$ , while for each random sample smaller than the estimated probability, a data point is randomly generated from the predicted distribution by the  $QRF^1$ . These data points

together estimate the full probability distribution of each record from first stage.

The models are trained and validated using the actual thunderstorm data obtained from a decade of data collection in Alabama. The studied area is divided into grid cells and all the data and predictions are produced per grid cell. Validating my models through holdout analysis, I demonstrate that my approach offers more accurate point and probabilistic predictions compared to traditional approaches. Comparing with the traditional two-stage modeling approach [74], the results of holdout analysis indicate that the proposed two-stage framework improves the accuracy of the point estimates. It is also found that applying cost-sensitive learning techniques in the first-stage results in not only more precise and computationally efficient point predictions, but also higher accuracy in probabilistic predictions. More accurate predictions produced by my modeling framework help utility companies make better decisions for post-storm restoration. The probabilistic predictions help them incorporate the existing uncertainty in the predictions in their decision making process.

To further improve the accuracy of the power-outage predictions, specially in much more zero-inflated data sets than the one used in this study, more research on the zero-inflation property is required. Moreover, due to recent technological and methodological progresses made in the data collection field, researchers are able to collect and store power-outage data more quickly than the past. Furthermore, the power system variables, weather conditions, and other parameters affecting the outages are changing over time. Therefore, future research is needed to further develop outage forecasting models to better adapt to changing conditions and data over time.

## CHAPTER III

# Adaptive Two-stage Bayesian Model Averaging for Estimating the Impact of Hazards on Power System Service <sup>1</sup>

### 3.1 Introduction

#### 3.1.1 Research Motivation

Weather events have the potential to cause disruptions in the electric power grid and result in power outages lasting from a few hours up to multiple days in the United States [74, 82, 117, 135]. As electrical power grids are highly interconnected with other critical infrastructures, a blackout may result in widespread economic and non-economic losses [26, 78]. A congressional research service study done in 2012 estimates the inflation-adjusted cost of weather-related outages at \$25 to \$55 billion annually, though there is a large annual variability [9, 26]. The frequency of such weather events has also been increasing over the last 30 years. It has been predicted that both the number and severity of them will increase due to global warming and climate changes [26]. Generally, the impacts of weather events on the electric power systems cannot be entirely prevented. However, by being better prepared, utility companies can restore service in a shorter time, and so reduce their costs [117]. Accurately estimating the number of power outages prior to a forecast event can help utility companies be better prepared and restore outages more quickly. Accordingly, in the recent years, more utility companies have started to build outage prediction models for forecasting storm impacts.

In the last two decades, a wide range of modeling techniques from parametric statistical models to nonparametric machine learning methods have been proposed for outage prediction

---

<sup>1</sup>Submitted to European Journal of Operational Research as Kabir, E., Guikema, S.D., and Quiring, S.M, Adaptive Two-stage Bayesian Model Averaging for Estimating the Impact of Hazards on Power System Service.

modeling (OPM). In almost all of these studies, several candidate models are chosen and trained on the collated power outage data. The models are then compared based on some accuracy criterion, typically using hold-out testing, and the best one is selected. Finally, this model is introduced as the one describing the process of weather-induced power outages, and all the inferences about the parameters and the variables are done based on this selected model. Despite their widespread use, however, there are a number of limitations with this modeling approach.

First, the uncertainty existing in the model is ignored. In these studies, one model is selected as the best one and all others are ignored even though the selected model may be only slightly more accurate than others. Parameter estimation, variable importance analysis and other inferential studies are all done based on the selected model ignoring the uncertainties existing in this model being the most accurate model in a given future scenario. Not conditioning the inferences on the selected model gives the decision maker a false impression of precision and hides the existing uncertainty in the model, which results in decision making that is not well-informed. Second, the current data may not be sufficient to select the best model explaining the data generating process and an observation of new data may lead to the selection of a completely different best model. This may yield sudden changes in estimates and the inclusion of new data may require revisiting a model that was previously rejected. This can appear to be erratic to model users. Third, in OPM, we care most about the predictive accuracy and the stability of the predictions rather than identifying the true model. Thus, we should look for a procedure that produces both accurate and stable predictions.

Despite over a decade of research in OPM, there still exists some amount of uncertainty in the forecasts which may arise from different sources such as the input data or the model choice. It is not enough to model the outages with a single output and instead, we should present the uncertainties that exist in the predictions as well. Based on conversations with utility storm personnel, in order to have a better recovery process and restoration planning, they need probabilistic models accurately estimating power outages in terms of probability distributions. However, most developed models estimate power outages as a single value rather than a probability distribution, and so, further research should be conducted in this area.

Furthermore, all developed models are based on traditional statistical or machine learning models and no adaptive learning methods have been developed for OPM. Adaptive prediction models can allow utility companies achieve better outage predictions while investing less time, effort and resources. This is because an adaptive model learns from the updates made to the input and output data. Therefore, the OPM literature needs models addressing this

adaptivity issue. Finally, almost all the models in the literature are developed for a single type of weather event, such as a hurricane (e.g., [74]) or a thunderstorm (e.g., [53]). There are relatively few studies addressing multiple types of weather events (e.g., [44]), but even in these studies, separate models are developed for different types of weather events. Thus, the power outage prediction literature lacks a single "all weather" model estimating outages for many different types of weather events simultaneously.

This chapter is motivated by the above-mentioned research gaps. It develops a new adaptive statistical learning approach based on Bayesian Model Averaging (BMA) to form an ensemble model estimating distributions of daily customer interruptions. I call this model Adaptive Two-stage Bayesian Model Averaging (AT-BMA).

### 3.1.2 My Contributions to the Literature

The contributions of this study are two-fold, as discussed below.

From a methodological perspective, I propose a new adaptive two-stage BMA algorithm. This proposed algorithm differs from a classical BMA approach in three aspects. First, unlike the traditional BMA approach, I formulate the weights of base learners (models) as an online multinomial logit model of the features. The posterior distribution of this model's parameters are approximated by using a Laplace approximation method. Then, I deploy a stochastic gradient ascent approach to update parameters of the posterior distribution as new data are observed. This helps not only tackle the challenge of likelihood estimation, but also provides a means to consider a variety of statistical and machine learning models such as quantile regression forest (QRF) as the algorithm's base learners. In other words, since the base learners' weights are affected by the number of times a model makes the best prediction among all considered models and not the distribution of error of each model, we can choose any type of statistical model for making the predictions and are not limited to specific parametric statistical models. Furthermore, in this approach, the weight of each base learner for a newly observed data point is different and based on the features of this record. This idea significantly extends the practicality of the BMA for various applications dealing with complex datasets.

Second, I have extended the BMA approach for the case of having multiple clusters of data instead of one comprehensive dataset. This makes my algorithm capable of modeling more complex data because we can initially divide the dataset into smaller clusters where data in each cluster are more similar to each other. Moreover, this leads to higher accuracy in the predictions. Third, in the classical BMA approach, only the weights of the base learners get updated as new data are collated and the base learners themselves are not updated. However, in my algorithm, I assign a training set to each base learner and newly collected

data are added to the training set of the model making the best prediction for that record. Periodically, each base learner is retrained on its own training set to get updated.

From an application perspective, although BMA has been employed successfully for prediction tasks in other disciplines, this is the first application of the BMA approach for OPM. Applying BMA for modeling weather-induced customer interruptions brings several advantages including: (i) formulating and representing model uncertainty, (ii) the ability to be updated as more data are accumulated, and (iii) resulting in optimal predictions under several loss functions, such as logarithmic, squared error loss, and continuous ranked probability score (CRPS) [106]. Finally, I develop a single algorithm for predicting daily customer interruptions over a large area covering multiple U.S. states and a wide variety of weather conditions. Such comprehensive model has never been developed in any of the past studies and can help utility companies make better restoration decisions.

### 3.1.3 Chapter Organization

This chapter is organized as follows. Section 3.2 provides a literature review of power outage predictive modeling. In Section 3.3, after describing the classical BMA approach, I introduce my proposed algorithm. Section 3.4 presents the case study. In this section I initially describe the large dataset used for validating my algorithm, and then provide the computational results from using the proposed method for modeling the daily customer interruptions. Finally, Section 3.5 concludes the chapter.

## 3.2 Literature Review: Weather-induced Outage Prediction Modeling

**Parametric statistical models:** A wide range of studies have been done in power outage prediction, beginning with the work of Davidson et al. [19]. Using a quantitative investigation of the performance of two electric power distribution systems, they showed that most tropical cyclone-related outages are caused by fallen trees on overhead power lines, and gust wind speed is a necessary, but not sufficient predictor of damage. Later, Liu et al. [68] utilized a negative binomial generalized linear model (GLM) to predict hurricane-related outages in the Carolinas. Han et al. [42] used the same model type in combination with principal component analysis, but with only variables that can be well measured prior to landfall, as opposed to the hurricane-indicator variables used by Liu et al. [68], to estimate the spatial distribution of hurricane outages in Gulf Coast region. Similarly, Zhou et al. [136] used a Poisson GLM and a Bayesian network to predict the yearly weather-related failure events on overhead lines. Comparing these two models, they found the Bayesian network

model preferable due to being (1) more informative than Poisson regression, (2) easier to implement, and (3) capable of getting updated with additional data.

Later, Liu et al. [66] used accelerated failure time models to predict outage duration from hurricanes and ice storms. One main limitation of their model was not incorporating key factors such as tree-trimming covariates or other features describing restoration resources (e.g., number of repair crews). Accordingly, Guikema et al. [36] developed a Poisson generalized linear mixed model (GLMM) to estimate the impacts of tree trimming on electric power system outages under normal operating conditions. Their results revealed that (1) increasing tree trimming frequency significantly decreases the number of resultant outages, and (2) the Poisson GLMM outperforms the negative binomial GLM for modeling the power outage data for one operating company. Liu et al. [67] also developed a GLMM using outage data from three large East coast utility companies for multiple hurricanes and ice storms to predict the spatial distribution of outages. However, their model still suffers from including the hurricane and ice-storm indicator variables because the models cannot be used for new storms. Building on Liu et al. [68, 67], Han et al. [41] developed a Poisson generalized additive model (GAM) in which the hurricane-indicator variable is replaced with physically measurable variables. The authors noted that the GAM could capture nonlinearity in the data, and overcome the over-prediction problems related to the negative binomial GLM. Having only physically measurable variables, they also could substantially improve the practical usefulness of the model.

**Non-parametric and Ensemble models:** One of the first uses of nonparametric models for hurricane outage prediction was conducted by Guikema et al. [38]. They developed two non-parametric models including classification and regression trees (CART) and Bayesian additive regression splines (BART). Comparing these two models with the previously introduced GAM and GLM showed that non-parametric approaches outperform the parametric ones in terms of predictive accuracy. Using a CART model, Quiring et al. [91] showed that the inclusion of certain soil and topographic variables significantly improves the predictive accuracy. Later, models based on ensembles of trees (explained in [51]) gained more popularity and Kankanala et al. [58] proposed an ensemble learning approach based on a boosting algorithm to estimate wind and lightning-related outages. Their results clearly showed that their proposed model estimates outages with greater accuracy than other models like neural networks and a mixture of experts. Shortly thereafter, Nateghi et al. [82] developed an ensemble model based on the method of random forest (RF) to estimate hurricane power outages for two states in the Gulf Coast region. They showed that their RF model yields much more accurate predictions using a significantly reduced number of predictors as compared with the Han et al. [42] and Liu et al. [68]. Wanik et al. [121] showed that an

ensemble model of RF, boosting tree and decision tree obtained the highest accuracy in terms of predicting the spatial distribution of outages per 2km grid cells.

To support wider emergency response planning, Guikema et al. [39] developed a Spatially-Generalized Hurricane Outage Prediction Model (SGHOPM) based on the RF method. This model was developed for the entire U.S. coastline. Using a RF coupled with a quantile regression forest (QRF) model, Tonn et al. [112] conducted a longitudinal study of outages at the zip code level to gain insight into the causal drivers of power outages during hurricanes. In another study, Wanik et al. [122] developed a RF to improve the OPM over eastern Connecticut for hurricanes. They incorporated tall vegetation that could come in contact with overhead distribution power lines, along with other vegetation management and infrastructure data. He et al. [44] conducted another study to compare two nonparametric tree-based models, QRF and BART, in terms of both power outage point estimates and prediction intervals for different types of weather events including hurricanes, blizzards, and thunderstorms. They found that BART produces more accurate point estimates than QRF, while QRF provides better prediction intervals for high spatial resolutions, but it does not aggregate well for coarser resolutions.

**Dealing with zero inflation:** Power-outage data, especially at high spatial and temporal resolutions, are zero-inflated. That is, the majority of locations experience no outages and so, the response variable has many zeros. This issue can result in some challenges in predictive modeling such as bias and inaccuracy. To deal with these issues, Guikema and Quiring [37] proposed a two-stage modeling approach in which the first stage, a CART model, predicts whether at least one outage will occur in each location. If the CART model determines that an outage will occur, a Poisson GAM estimates the number of outages in that area. They showed that their two-stage model outperforms the classical zero-inflated Poisson GLM and zero-inflated negative binomial GLM for their power outage prediction problem. The SGHOPM was extended by McRoberts et al. [74] to the two-stage approach using RF in both stages. Later, the two-stage modeling technique was extended by Kabir et al. [53] for predicting thunderstorm-induced customer interruptions which are highly zero-inflated. In their approach, resampling and cost-sensitive learning are incorporated within the first-stage model, and the QRF is used in the second-stage to produce probabilistic prediction. In their approach, instead of producing point estimates, the full probability distributions of the number of customers interrupted in each grid-cell are estimated. They showed that incorporating both resampling and cost-sensitive learning techniques enhances the accuracy of predictions specifically for the non-zero class. They also showed that their approach outperforms the two-stage model developed by McRoberts et al. [74].

The OPM literature shows that a wide range of modeling techniques have been employed

in past studies. This highlights the model uncertainty issue that was discussed in section ?? . Furthermore, most OPMs are developed for a single weather event (e.g., [37, 39, 53, 74]). There are a few cases where separate models are developed for different weather events, but for each type of weather one model is selected to make a prediction (e.g., [44]). This approach is operationally challenging for the utility companies because it is not always clear what type of weather event it is and which model should be run. Accordingly, utility companies are looking for models that are not weather-event specific and can make predictions for all types of weather events. Finally, studying the literature shows the lack of an adaptive model able to learn from new data and keep the system updated. Thus, in this study, I address the above-mentioned gaps and develop an adaptive all-weather model.

### 3.3 Model

#### 3.3.1 Model Averaging

The selection of one particular model among a set of trained ones may lead to overconfident predictions and riskier decision making because it ignores the existent model uncertainty in favor of very particular distributions and assumptions on the model of choice. This modeling strategy, which is used by many researchers, is called model selection. To deal with model uncertainty, an alternative approach that has attracted increasing attention is model averaging. In this approach, I take into account all the models existing in the model space and the prediction is averaged over all these models using weights. Model averaging has two main strands: Bayesian model averaging and frequentist model averaging.

Bayesian Model Averaging (BMA) introduced by Leamer [64] and later used in Min and Zellner [77], and Raftery et al. [92] is a direct consequence of the Bayes' theorem in a model uncertainty setting. BMA adds a layer to the hierarchical modeling in Bayesian inference by assuming a prior distribution over the set of all considered models, which describes the prior uncertainty over each model's capability to accurately describe the data. Let  $\mathcal{M} = (\mathcal{M}^1, \mathcal{M}^2, \dots, \mathcal{M}^K)$  denote the set of models under consideration. There is a probability mass function (prior) over all the models with values  $p(\mathcal{M}^k) \forall k = 1, \dots, K$ . By using Bayes' theorem, I derive posterior model probabilities given the observed data by:

$$p(\mathcal{M}^k|D) = \frac{p(D|\mathcal{M}^k)p(\mathcal{M}^k)}{\sum_{m=1}^K p(D|\mathcal{M}^m)p(\mathcal{M}^m)}$$

where  $D$  is the observed data, and  $p(D|\mathcal{M}^k)$  is the model's marginal likelihood or model evidence. Suppose we are interested in estimating a quantity of interest  $\Delta$ , such as a future observation or a model parameter, on the basis of the data. Then, its marginal posterior

distribution across all models is obtained by:

$$p(\Delta|D) = \sum_{k=1}^K p(\Delta|D, \mathcal{M}^k)p(\mathcal{M}^k|D)$$

This is an average of the posterior predictive distributions for  $\Delta$  over all the considered models, weighted by their corresponding posterior model probability. In the classical BMA, it is assumed that the model space contains the "true" data generating model, although it is unknown. In this situation, which is referred to as M-closed,  $p(\mathcal{M}^k|D)$  is the posterior probability that model  $\mathcal{M}^k$  is true. In BMA, there are priors for both models and model-specific parameters, and their specification is quite important for the final outcomes and the posterior model probabilities. In some simple settings, such as GLMs with conjugate priors, the model's marginal likelihood can be calculated analytically, but in general it does not have a closed form and must be approximated. Some well-known approximations are done through using Bayesian Information Criteria (BIC), Kullback Information Criterion (KIC), Bayes factor, or Akaike information criterion (AIC) [28]. These approximations might not be adequate for more complex models because they rely on many regularity conditions.

Classical BMA focuses on identification of the true model in an M-closed framework. However, this assumption is not realistic in many situations. Recently, several researchers have considered the M-open framework in which it is assumed that the true model does not exist in the model space. Under this assumption, models' posterior probabilities are determined using cross validation [131]. Several empirical and theoretical analyses of the performance of the BMA approach have been done in the literature. They show that BMA leads to better predictions with lower variance under a logarithmic scoring rule [72] rather than using a single model. This practice is also considered as a standard post-processing approach in order to make inference in the presence of multiple competing statistical models for producing probabilistic forecasts [93, 104]. In the frequentist model averaging (FMA), unlike BMA, no priors are considered on the candidate models and the outputs are point estimates. Model weights and parameters are totally determined by data. For an overview of FMA, see [28].

### 3.3.2 Proposed model: Adaptive two-stage BMA

Starting in the late 1980s, Bayesian modeling and inference gained attention in the reliability literature. Since then, advances in computational abilities have significantly contributed not only to an increase in their implementation but also in their use for solving decision-making problems [49, 130]. In this study, I present a new BMA approach to tackle

model uncertainty. I relax the classical assumption that the model space includes the true model. I treat the models as part of the action space and my objective is to combine their predictions with each other to achieve improved forecasts. In my approach, model weights represent how close they are to the true model and they are functions of the data. In the following, I explain this approach in detail.

### 3.3.2.1 Bayesian Model Averaging for Clustered Data

Initially, the data are divided into smaller clusters in such a way that the objects in the same cluster are more similar. I denote these clusters by  $\mathcal{C}^1, \mathcal{C}^2, \dots, \mathcal{C}^L$  where  $L$  is the number of clusters. Next, I assume that there are  $K$  candidate probabilistic base learners, denoted by  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K$ . Each of these  $K$  base learners has a separate training set represented by  $\Phi_1, \Phi_2, \dots, \Phi_K$ . The goal is to combine the predictions of these base learners such that the error in the final prediction is lower and more stable. To estimate the probability distribution of the response variable  $y$  given a newly observed covariate vector  $x$ , I condition on the cluster to which the covariate vector  $x$  belongs. Thus, we have the following probability distribution of the response variable  $y$  given observed covariate  $x$ .

$$p(y|x) = \sum_{l=1}^L p(y|x, \mathcal{C}^l) p(\mathcal{C}^l|x),$$

where  $p(\mathcal{C}^l|x)$  is the probability of covariate  $x$  belonging to the cluster  $\mathcal{C}^l$ . As an example, in a Gaussian Mixture Model (GMM),  $p(\mathcal{C}^l|x)$  is equal to probability density of a data point  $x$  in  $l^{th}$  multivariate Gaussian model divided by sum of probability density values obtained from all multivariate Gaussian models at point  $x$ . We can also think of K-means clustering as a GMM with fixed variance components.

The marginal posterior distribution of the response variable  $y$  across all  $K$  base learners is given by:

$$p(y|x, \mathcal{C}^l) = \sum_{k=1}^K p(y|x, \mathcal{C}^l, \mathcal{M}^k) p(\mathcal{M}^k|x, \mathcal{C}^l).$$

We shall refer to the above equation as Bayesian model averaging where  $p(y|x, \mathcal{C}^l, \mathcal{M}^k)$  is the predictive probability density function based on base learner  $\mathcal{M}^k$  given covariate vector  $x$  and knowing that  $x$  belongs to cluster  $\mathcal{C}^l$ . Here,  $p(\mathcal{M}^k|x, \mathcal{C}^l)$  denotes the posterior probability that the base learner  $\mathcal{M}^k$  is the most accurate model given  $x \in \mathcal{C}^l$  (i.e.  $\mathcal{M}^k$  makes the best forecast for covariate  $x$ ). In the following, I formulate the posterior probabilities as a multinomial logit model and show how it can be updated with each newly observed data. Figure 3.1 represents my algorithm graphically.

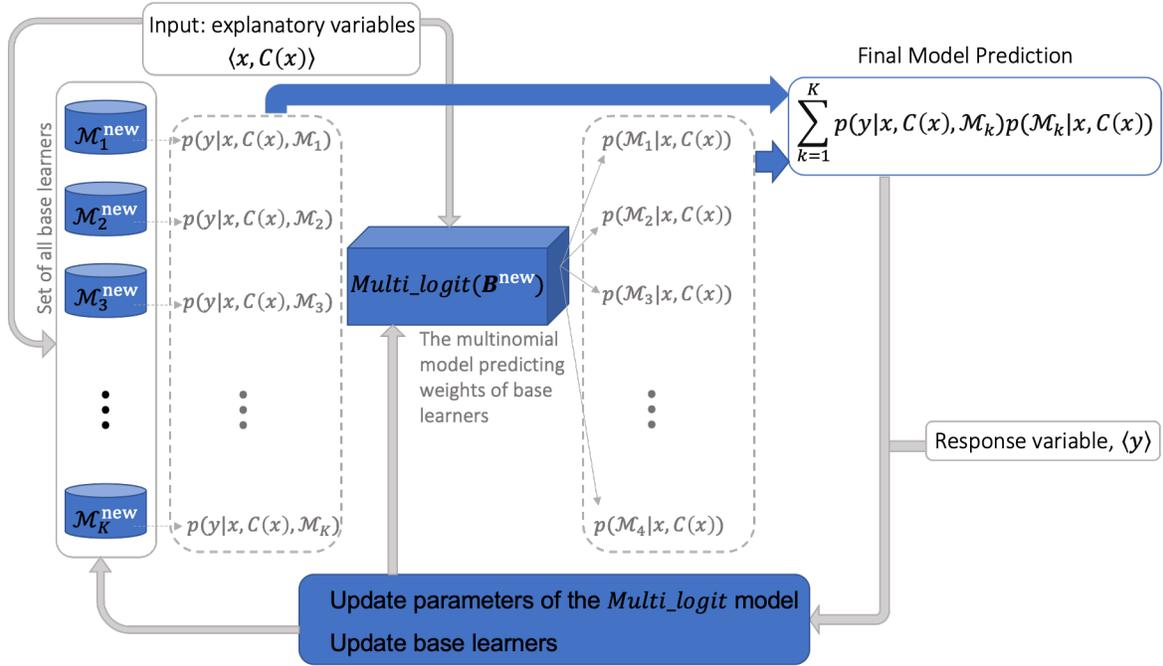


Figure 3.1: This figure summarizes my proposed adaptive two-stage Bayesian model averaging algorithm. There are  $K$  base learners represented by  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K$ . Each base learner independently predicts probability distribution of response variable  $y$  for each covariate vector  $x$  knowing the cluster this record belongs to,  $C(x)$ . The final probabilistic prediction for covariate vector  $x$  is a weighted average over all predicted probability distributions. The weights of models in the final prediction is a multinomial logit model of covariate  $x$ . Parameters of the multinomial logit model as well as all the base learners are updating over time using newly observed data.

### 3.3.2.2 Models' Posterior Probabilities as a Multinomial logit model

In my proposed algorithm, the performance of candidate models does not have to be the same for all records. I formulate the probability of models making the best prediction as a multivariate logit function of covariate vector  $x$ :

$$p(\mathcal{M}^k|x, \mathcal{C}^l) = \begin{cases} \frac{\exp(\beta_k^T x)}{Z_x} & \text{for } k = 1, \dots, K-1 \\ \frac{1}{Z_x} & \text{for } k = K \end{cases}$$

where  $Z_x = 1 + \sum_{k=1}^{K-1} \exp(\beta_k^T x)$  and  $\beta_k = [\beta_{k1}, \dots, \beta_{kd}]_{k=1, \dots, K-1}^T$  is a parameter vector corresponding to the  $k^{\text{th}}$  model. If  $\mathbf{B} = [\beta_1, \dots, \beta_{K-1}]$  represents the matrix of parameters, the posterior distribution of  $\mathbf{B}$  given observed data  $D$  is calculated as follows:

$$\begin{aligned} p(\mathbf{B}|D) &\propto \{p(D|\mathbf{B}) \cdot p(\mathbf{B})\} \\ \log p(\mathbf{B}|D) &\propto \{\log p(D|\mathbf{B}) + \log p(\mathbf{B})\} \end{aligned}$$

where  $D = \langle x_i, y_i \rangle_{i=1, \dots, n}$  is a sequence of  $n$  data points with  $x_i \in \mathbb{R}^d$  and  $y_i \in \{1, \dots, K\}$ ,  $p(D|\mathbf{B})$  is the likelihood of the data in the model with parameter matrix  $\mathbf{B}$  and it is computed as:

$$p(D|\mathbf{B}) = \prod_{i=1}^n \frac{\exp(\beta_{y_i}^T x_i)}{1 + \sum_{j=1}^{K-1} \exp(\beta_j^T x_i)},$$

where  $\beta_K$  is a  $d$  dimensional vector of zeros. Also,  $p(\mathbf{B})$  is the prior of the parameter matrix. I assume this prior is a univariate Gaussian with mean  $\mu_{kj}$  and variance  $\sigma_{kj}^2$  on each parameter  $\beta_{kj}$ . I also assume that the components of  $\mathbf{B}$  are independent and hence, the overall prior for  $\mathbf{B}$  is the product of the priors for its components:

$$p(\mathbf{B}) = \prod_{k=1}^{K-1} \prod_{j=1}^d \left( \frac{1}{\sqrt{2\pi}\sigma_{kj}} \exp\left(-\frac{(\beta_{kj} - \mu_{kj})^2}{2\sigma_{kj}^2}\right) \right).$$

Exact Bayesian inference for the posterior distribution of parameters of a multinomial logistic regression is intractable. I can either deploy analytic approximations to the posterior, or solutions based on Markov Chain Monte Carlo (MCMC) methodology [29]. The MCMC is usually computationally inefficient in terms of both time and space complexity. Therefore, in what follows, I consider the Laplace approximation method which approximates the posterior distribution with a Gaussian distribution. If I denote the log posterior probability of parameters with  $\Psi(\mathbf{B}) = \log p(D|\mathbf{B}) + \log p(\mathbf{B})$ , the second-order Taylor expansion of this

---

**Algorithm 3** Adaptive Two-stage Algorithm based on Bayesian Model Averaging
 

---

- 1: **Given:** Models  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K$ ; Sets  $\Phi_1, \Phi_2, \dots, \Phi_K$ ; and Clusters  $\mathcal{C}^1, \mathcal{C}^2, \dots, \mathcal{C}^L$
- 2: **Initialize:** Mean and variance (respectively  $\mu_{kj}$  and  $\sigma_{kj}^2, \forall k = 1, \dots, K - 1 \& \forall j = 1, \dots, d$ ) of Gaussian prior distributions defined on the parameters of the multinomial logit model.
- 3: **for**  $t = 1, 2, \dots$  **do**
- 4:   Get new batch of data, called  $D = \langle x_i, \cdot \rangle_{i=1, \dots, n}$ .
- 5:   **for** each record  $x_i \in D_t$ ;  $\forall i = 1, 2, \dots, n$  **do**
- 6:     Estimate probability of the record belonging to each cluster,  $p(\mathcal{C}^l | x_i) \forall l = 1, \dots, L$
- 7:     Using each base learner, make prediction for the probability distribution of the response variable. The produced probabilistic prediction is shown by  $p(y_i | x_i, \mathcal{C}^l, \mathcal{M}_k) \forall k = 1, \dots, K$
- 8:     Set  $\beta_k = [\mu_{k1}, \dots, \mu_{kd}]^T$  for  $k = 1, \dots, K - 1$  and  $Z_{x_i} = 1 + \sum_{k=1}^{K-1} \exp(\beta_k^T x_i)$ . Estimate probability of each base learner making the best prediction for covariate vector  $x_i$  as follows:

$$p(\mathcal{M}^k | x_i, \mathcal{C}^l) = \begin{cases} \frac{\exp(\beta_k^T x_i)}{Z_{x_i}} & \text{for } k = 1, \dots, K - 1 \\ \frac{1}{Z_{x_i}} & \text{for } k = K \end{cases}$$

- 9:     Make the final probabilistic prediction for  $x_i$ , shown by  $p(y_i | x_i)$ , according to the following formula:

$$p(y_i | x_i) = \sum_{l=1}^L p(y_i | x_i, \mathcal{C}^l) p(\mathcal{C}^l | x_i)$$

$$p(y_i | x_i, \mathcal{C}^l) = \sum_{k=1}^K p(y_i | x_i, \mathcal{C}^l, \mathcal{M}_k) p(\mathcal{M}_k | x_i, \mathcal{C}^l)$$

- 10:     Observe the actual response variable  $y_i$  for record  $x_i$  and set  $b_i^k$  equal to 1 if  $k^{th}$  model makes the best prediction for  $x_i$ , and 0 otherwise.
  - 11:     Add the new record  $\langle x_i, y_i \rangle$  to set  $\Phi_j$  where  $j = \operatorname{argmax}_k b_i^k$
  - 12:     Update  $\mu_{kj}$  and  $\sigma_{kj}^2, \forall k = 1, \dots, K - 1 \& \forall j = 1, \dots, d$  according to Algorithm 2.
  - 13:     Update the models: retrain each model on the updated corresponding training set
- 

function at point  $\hat{\mathbf{B}}$  is written as:

$$\Psi(\mathbf{B}) \approx \Psi(\hat{\mathbf{B}}) + (\mathbf{B} - \hat{\mathbf{B}}) \Psi'(\hat{\mathbf{B}}) + \frac{1}{2} (\mathbf{B} - \hat{\mathbf{B}})^2 \Psi''(\hat{\mathbf{B}}).$$

By choosing  $\hat{\mathbf{B}}$  at the peak (mode) of the  $\Psi(\cdot)$ , where the derivative is zero, the posterior distribution of parameters can be approximated with a Gaussian centered at the mode:

$$\mathbf{B} | D \sim \text{Gaussian}(\hat{\mathbf{B}}, -\Psi''(\hat{\mathbf{B}}))$$

The problem is thus reduced to find  $\hat{\mathbf{B}}$  such that:

$$\begin{aligned} \hat{\mathbf{B}} &= \operatorname{argmax}_{\beta} \Psi(\mathbf{B}) \\ &= \operatorname{argmax}_{\beta} \{ \log p(D | \mathbf{B}) + \log p(\mathbf{B}) \} \end{aligned}$$

To obtain  $\hat{\mathbf{B}}$ , I deploy a stochastic gradient ascent algorithm. Details of this optimization

---

**Algorithm 4** Stochastic Gradient Ascent Algorithm for finding the Optimal Solution of  $\Psi(\mathbf{B})$  Function

---

- 1: **Given:** Batch of Data  $D = \langle x_i, y_i \rangle_{i=1, \dots, n}$ ; Dimension of feature space and response variable ( $d$  and  $K$  respectively); Mean and variance of prior distributions of model parameters  $(\mu_{kj}, \sigma_{kj}^2) \forall k = 1, \dots, K - 1$  &  $j = 1, \dots, d$
  - 2: **Initialize:** Maximum number of training iteration,  $T$ ; Minimum relative error improvement per iteration,  $\epsilon \in \mathbb{R}^+$ ; Initial learning rate,  $\eta_0$ ; Annealing rate,  $\delta \in \mathbb{R}^+$
  - 3: **for**  $t = 1, 2, \dots, T$  **do**
  - 4:      $\eta_t = \frac{\eta_0}{1+t/\delta}$
  - 5:     **for**  $i = 1, 2, \dots, n$  **do**
  - 6:          $Z \leftarrow 1 + \sum_{k=1}^{K-1} \exp(\beta_k^T x_i)$
  - 7:         **for**  $k = 1, 2, \dots, K - 1$  **do**
  - 8:              $p(k|x_i, \mathbf{B}) \leftarrow \exp(\beta_k^T x_i) / Z$
  - 9:              $\beta_k \leftarrow \beta_k + \eta_t (\frac{1}{n} \nabla_k \Psi(\mathbf{B}))$
  - 10:      $l_t = \Psi(\mathbf{B})$
  - 11:     **if**  $\frac{|l_t - l_{t-1}|}{|l_t|} < \epsilon$  **then**
  - 12:         Return  $\mathbf{B}$
- 

methodology are given by Algorithm 4, where elements of the  $(K - 1) * d$  dimensional matrix  $\nabla \Psi(\mathbf{B})$  is calculated as:

$$\nabla_{k,j} \Psi(\mathbf{B}) = \sum_{i=1}^n \left( x_{ij} (I(k = y_i) - p(k|x_i, \mathbf{B})) - \frac{\beta_{kj} - \mu_{kj}}{n\sigma_{kj}^2} \right),$$

and  $\nabla_k \Psi(\mathbf{B})$  is the  $k^{th}$  row of the gradient matrix. Using the stochastic gradient ascent (Algorithm 4), I am able to estimate the mean and variance of the posterior distributions of parameters of the multinomial logit (ML) model. In Algorithm 2, in each iteration of receiving new data and updating the parameters of the ML model, the previous values of parameters are used as the prior for the next iteration. For every instance, the ML model predicts the probability that each base learner makes the best prediction. These predicted probabilities are the weights of base learners' in the BMA model.

### 3.3.2.3 Updating the models

In my proposed adaptive two-stage algorithm, I assume that each base learner  $\mathcal{M}^k; k = 1, \dots, K$  has a training set denoted by  $\Phi_k; k = 1, \dots, K$ . Each of these training sets is updated when observing new data. Every time I receive a new record, this record is added to the training set of the base learner which makes the best prediction for that record. I then periodically retrain each base learner on its own updated training set. See Algorithm 3 to find my approach step by step.

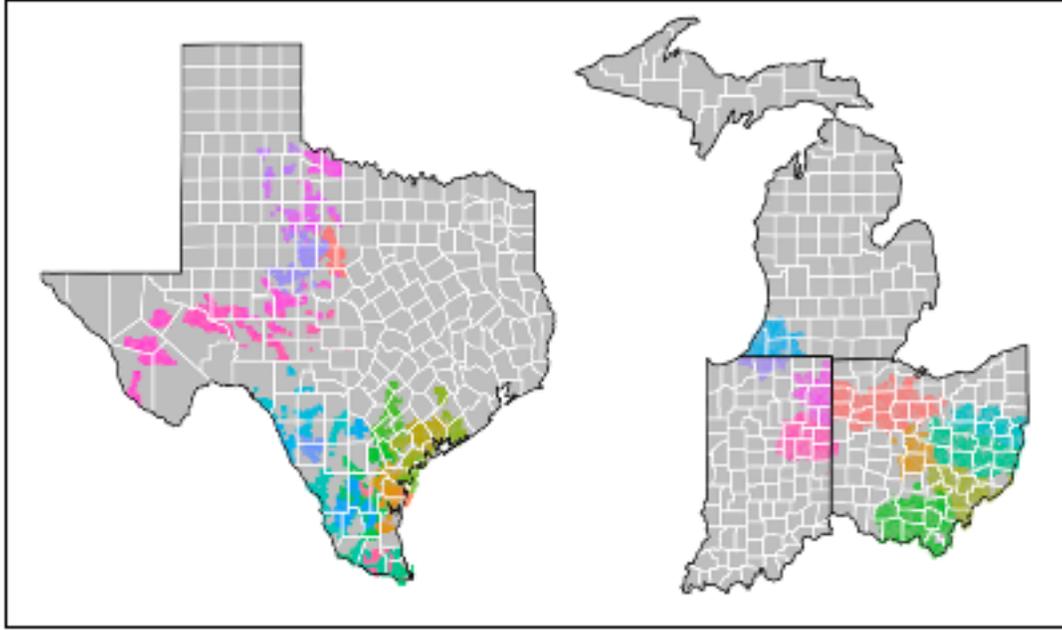


Figure 3.2: Spatial extent of the service territory for AEP Ohio, I&M Power and AEP Texas. The total service territory is divided into 46 subareas that are defined by the utility company and shown in different colors.

### 3.4 Case Study: Predicting The Number Of Customers Interrupted

#### 3.4.1 Data Description

My case study uses the daily number of customers interrupted from 2012 to 2018 in the service territory of a major utility serving Michigan, Indiana, Ohio and Texas. These data were provided by American Electric Power (AEP). The colored areas in Figure 5.1 show the spatial extent of the service territory. This service territory is divided into 46 subareas that are defined by the utility company and I model at these subareas. Every reported outage was recorded with an address and the number of affected customers by this outage. I geolocated and aggregated all outages that occurred within the same subarea. Therefore, I define the response variable as the total number of customers interrupted in each subarea.

The covariates used in this study are divided into two main categories, static and dynamic. The first group is related to the power system and includes time-invariant covariates, and the second one is related to weather, precipitation and soil moisture that changes every day. All the covariates and response variables are aggregated to the level of subareas. Each group of variables is explained below.

**Static covariates:** To characterize the power system, I include the number of poles,

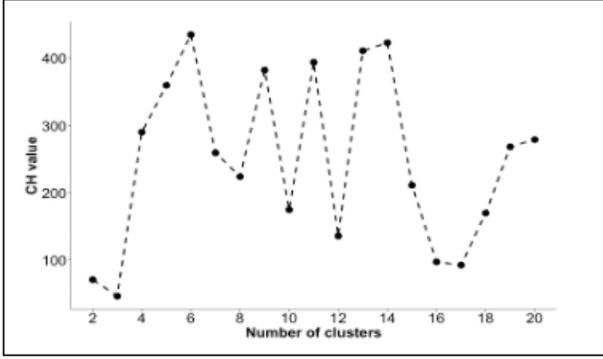


Figure 3.3: CH factor is estimated for different number of clusters. Having six clusters results in the optimum CH value.

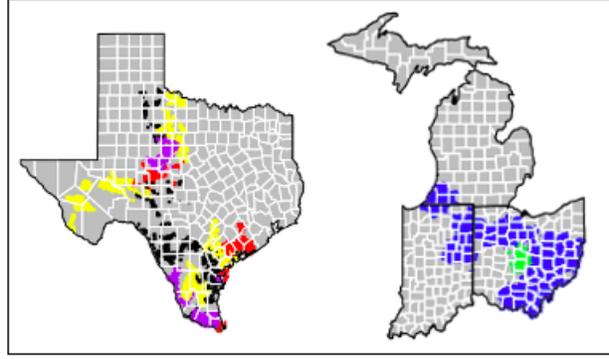


Figure 3.4: Data is clustered based on the static variables, and different colors show the resultant six clusters.

switches, overhead and under-ground transformers, reclosers, fuses, and total length of overhead and underground lines in each subarea. These variables provide a measure of the extent of power system exposure to weather events. The number of customers in each subarea is another variable included in my model. The static variables are used in the clustering in order to divide the whole dataset into smaller ones. The Calinski-Harabasz (CH) factor is deployed as a measure to choose the optimal number of clusters. It is defined as  $\frac{\text{betweenSS}/k-1}{\text{withinSS}/n-k}$ , where  $k$  is the number of clusters, and  $n$  is the dataset length. BetweenSS is the average of distances between cluster centers, and withinSS is the average of distances from each record to the center of its own cluster. Ideally, I would like to have a clustering that has the properties of internal cohesion and external separation. Thus, I look for the  $k$  maximizing the CH factor. Based on this, the optimal number of cluster is chosen to be 6 (see Figure 3.3). Figure 3.4 illustrates the spatial distribution of these six clusters.

**Dynamic covariates:** The dynamic covariates used in this study are soil moisture, historic precipitation levels, and weather forecast variables. Soil moisture and precipitation data are derived from the North America Land Data Assimilation System. Soil moisture is extracted at three depth levels including 0-10 cm, 10-40 cm, and 40-100 cm. The values of total water volume are converted to volumetric water content and then mapped to an empirical cumulative distribution function (CDF). Soil moisture CDFs for the three mentioned depth levels are used in the model. The standardized precipitation index (SPI) is a measure of precipitation deviations from normal conditions. SPI is also estimated for different durations of 1, 3, 6, and 12 months where an  $n$ -month SPI is a measure of the deviation of precipitation from the long-term average in the  $n$  months prior. The weather data were obtained from the National Digital Forecast Database. They include temperature, maximum instantaneous wind gust (m/sec) speed, probability of a tornado, hail, and damaging

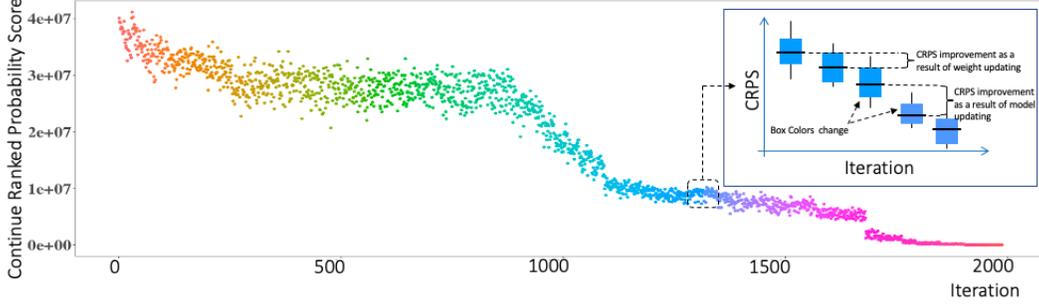


Figure 3.5: Decrease of the CRPS error of the BMA model as a function of iterations.

thunderstorm, snow amount and 24 hour quantitative precipitation forecast (QPF).

### 3.4.2 Evaluation metric

The continuous ranked probability score (CRPS) is a proper scoring rule addressing two important aspects of probabilistic forecast which are calibration and sharpness [34]. These two aspects help to ensure that the forecast is accurate and the predicted distribution is concentrated. CRPS compares the cumulative distribution function of the prediction to that of the observed data. It is defined as

$$CRPS(F, y_0) = \int_{-\infty}^{\infty} (F(y) - \mathbb{1}\{y \geq y_0\})^2 dy,$$

where  $F$  is the predicted CDF for the response value  $y_o$ .

### 3.4.3 Computational results and analysis

To test and validate my proposed algorithm, I use daily customer interruptions data. My dataset includes the data from 1988 days of customer interruptions in 46 subareas in Indiana, Michigan, Ohio and Texas. To test the predictive accuracy of my model, I use holdout testing. In every hold-out test, I leave one month of data out. I then train the model on the remaining 71 month data, and test it against the held-out data.

As we can see in Algorithm 3, to train the AT-BMA model, we need a number of models referred as the base learners  $(\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K)$ , as well as their corresponding training sets  $(\Phi_1, \Phi_2, \dots, \Phi_K)$ . I choose four probabilistic model types including Bayesian additive regression tree (BART), Bayesian classification and regression tree (BCART), QRF, and Bayesian linear regression model (BLM) in order to construct the initial base learners. Then, based on the spatial clustering distribution shown in Figure 3.4, I divide the first month of data from the training set into six clusters. Next, I train the above-mentioned models on each of

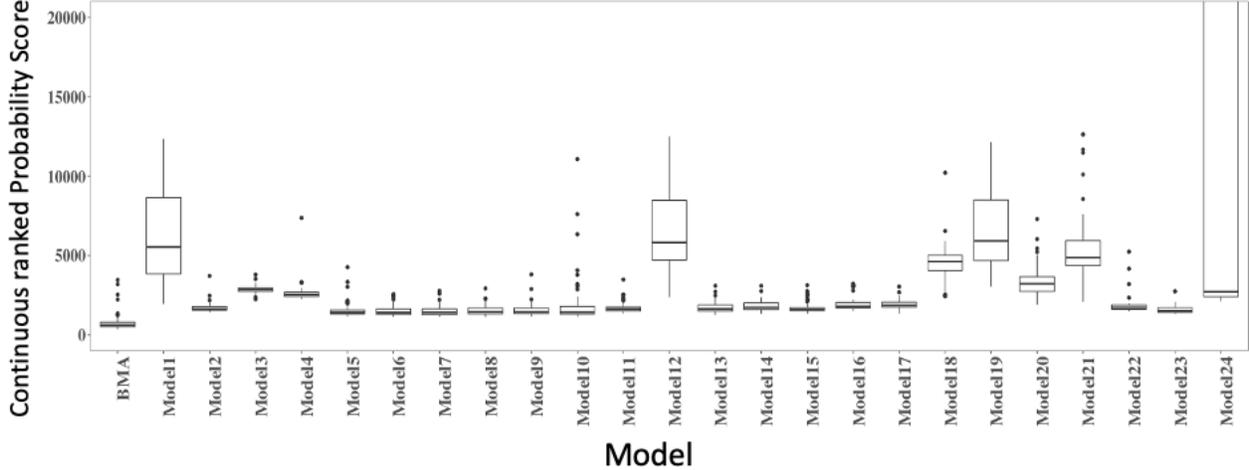


Figure 3.6: Boxplot comparison of the CRPS errors of the proposed BMA model with its base learners.

these clusters of data using all 55 dynamic covariates described in section 4.3. Therefore, I construct  $K = 24$  initial base learners that are not similar. The data used for each model being trained are also considered as the base training sets ( $\Phi_i; i = 1, \dots, 24$ ). In the first iteration and before any information is obtained from the data, I set the prior mean to be zero and the prior standard deviation to 10 (i.e.,  $\mu_{kj} = 0$  and  $\sigma_{kj}^2 = 100$  for all  $k$  and  $j$ ) for all the parameters of the multinomial logit model which is formulating the weights of base learners.

In every iteration of the algorithm, the data from one day including the number of customers interrupted in all 46 subareas are observed by the algorithm (i.e.,  $t = 1, \dots, 1988$  and  $n = 46$  in Algorithm 3). For each newly observed data point  $x_i, i = 1, \dots, 46$ , the probability of the record belonging to cluster  $C^l$  is estimated as the distance of  $x_i$  to the center of cluster  $C^l$  divided by sum for distances of  $x_i$  to the centers of all the clusters. Each base learner produces a probabilistic prediction for each record. Then, the multinomial logit model estimates the probability of each base learner making the best prediction for record  $x_i$ . The final probabilistic prediction is obtained based on step 9 in Algorithm 3.

The parameters of the multinomial logit model are updated according to Algorithm 4 as any new record of data is observed. In every iteration of  $t = 1, \dots, 1988$ , after the true number of customers interrupted value (response variable), is observed, I calculate the CRPS value for each forecast obtained by the base learners. Then, for each record, I find the model which made the best prediction (i.e., had the least CRPS value). This information creates a new dataset  $D = \langle x_i, y_i \rangle_{i=1, \dots, 46}$ , where  $y_i \in \{1, 2, \dots, 24\}$  and  $x_i$  is a  $d = 21$  dimensional vector of covariates that are shown in the first row of Table 3.1 (the cluster covariate is a categorical variable with 6 levels and so, I model it with five binary variables). To implement

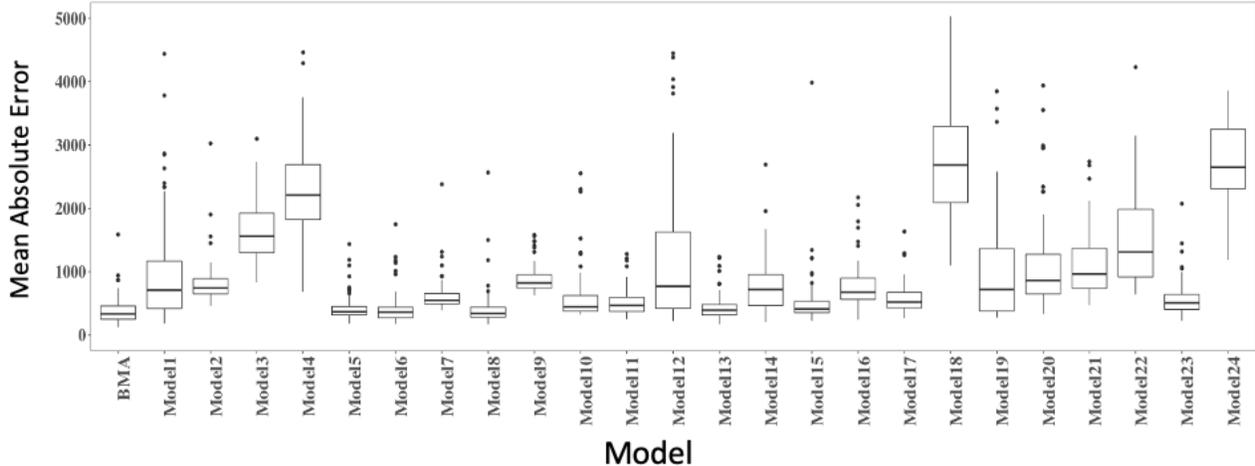


Figure 3.7: Comparison of the MAE error of the proposed BMA model with its base learners.

Algorithm 4 and update parameters of the multinomial logit model, I set  $T = 100$ ,  $\epsilon = 0.01$ ,  $\eta_0 = 1$  and  $\delta = 1$ . Each newly observed record is added to the training set of model making the best prediction for that record. Then, after observing data of one month, I retrain the base learners on their updated training sets.

As discussed previously, I held out data of one month and trained the algorithm on the training set step by step. Every day, after the model is updated using newly observed data, I make prediction for the test set. Therefore, in total, I make predictions for the test set 1988 times. I also repeat this process by choosing every one month of data as the test set. So, the process is repeated for 72 times. This analysis results in Figure 3.5. In this figure, due to the limited space, instead of a boxplot of the 72 CRPS values in each iteration, I only represent the median as a single point. Furthermore, data points related to each month are shown with different colors. Figure 3.5 indicates that the AT-BMA algorithm makes more and more accurate predictions for the test set as more data is revealed and used. I also find that the improvement in the model is a result of both model updating (change of color), and model weights. We see that 90% of the improvements in the CRPS value are due to model updating, and the remaining 10% are due to weight updating. Therefore, both practices are successful in improving the model performance.

I also compare the AT-BMA model with the predictive performance of its base learners. Figures 3.6 and 3.7 illustrate that the out-of-sample CRPS and MAE of the AT-BMA are less than the CRPS and MAE of all base learners individually. This shows that the model averaging approach results in obtaining more accurate predictions. The lower CRPS value obtained by the BMA model is an expected result because I choose weights of base learners such that the CRPS value of the combined prediction is minimized. However, Figure 3.7

Table 3.1: Average coefficients of the multinomial logit model in the AT-BMA algorithm are shown in this table. Each row of the table indicates the average of coefficient values for each model. The multinomial logit model is built of 21 variables including an intercept, 15 continuous weather-related variables, and 5 binary variables associated with the cluster covariate. The larger each variable coefficient is, the better prediction the model makes for records with larger values of that variable.

Model	Intercept	Mean wind gust speed	Max wind gust speed	Mean probability of damaging thunderstorm	Max probability of damaging thunderstorm	Mean probability of hail	Max probability of hail	Mean probability of tornadoes	Max probability of tornadoes	Mean snow amount	Max snow amount	Mean QPF	Max QPF	Mean temperature	Min temperature	Max temperature	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1	-0.75	0.08	-0.07	-0.36	0.06	<b>0.27</b>	<b>0</b>	<b>1.17</b>	<b>0.79</b>	-0.57	-0.35	0.01	-0.01	-0.05	0.04	0.03	0.09	-1.19	0.27	-2.63	-1.75
2	2.65	-0.03	-0.03	0	-0.12	0.05	0	-0.04	0.08	0.09	-0.01	-0.02	-0.03	0.06	-0.01	-0.05	<b>1.2</b>	-0.78	0.88	-2.89	-1.19
3	0.05	0.1	-0.06	0.17	-0.25	-0.39	0.16	0	0.01	-0.07	-0.05	-0.17	0.07	0.04	0.01	-0.06	0.38	<b>1.05</b>	0.33	0.47	-0.55
4	2.84	0.07	-0.07	-0.07	-0.07	0.2	-0.04	-0.47	0.39	<b>1.96</b>	<b>1</b>	-0.31	-0.02	0.09	-0.03	-0.06	0.84	-0.26	<b>1.14</b>	-3.59	-0.97
5	-1.08	0.04	-0.06	-0.31	0.16	0.38	-0.18	-0.26	-0.52	-0.2	-0.1	0.06	-0.02	0.03	-0.04	0.01	1.03	0.85	0.59	<b>1.35</b>	-0.04
6	-1.23	0.04	-0.07	-0.35	0.19	0.38	-0.19	-0.19	-0.08	-0.44	-0.26	0	-0.02	-0.03	0.03	0.01	0.14	-0.75	-0.09	-1.18	-1.01
7	-0.2	0.21	-0.1	0.37	-0.22	-0.36	0.12	0.09	0.15	-0.12	-0.06	-0.05	0.01	0.05	-0.04	-0.01	0.25	-0.38	0.46	0.77	-0.61
8	-0.22	0.21	-0.16	-0.15	0.12	0.26	-0.04	-0.25	-0.07	-0.06	-0.03	0.1	0.05	-0.04	-0.03	0.07	-0.06	0.67	-0.2	0.62	-0.4
9	-0.11	<b>0.23</b>	<b>0.17</b>	-0.35	-0.4	-0.54	-0.81	-0.22	-0.49	-0.05	-0.02	-0.04	0.07	0.15	0.11	-0.27	0.65	-0.52	-0.64	0.05	<b>0.72</b>
10	-0.13	-0.15	0.09	0.13	-0.2	0.02	0.15	-0.35	0.11	-0.07	-0.04	0.16	-0.07	-0.02	-0.04	0.05	-0.04	<b>1.43</b>	-0.5	0.18	-0.89
11	0.03	0.04	0	0.03	-0.26	-0.03	-0.06	0.47	<b>0.83</b>	0.01	0	0.83	-0.34	0.11	-0.12	0	-0.27	0.56	-0.46	-0.71	<b>0.33</b>
12	-0.09	0.6	-0.19	<b>0.23</b>	<b>0.21</b>	-0.5	0.03	-0.14	-0.34	-0.02	-0.01	-0.36	0.03	-0.28	0.21	0.04	-0.24	-0.51	-0.38	<b>1.18</b>	-0.31
13	-0.36	0.06	-0.05	-0.41	0.17	0.4	-0.14	0.31	-0.37	-0.09	-0.09	-0.21	0.04	-0.05	0.08	-0.02	-0.62	-0.07	0.09	-0.25	-0.98
14	0.12	0.01	0	0.04	-0.11	0.14	-0.07	0	-0.64	0.09	0.11	-0.21	0.08	0.06	-0.04	-0.01	-0.34	-0.17	0.03	-1.83	-1.07
15	-0.1	0.09	-0.07	0.1	-0.07	-0.06	0.03	0.01	-0.12	-0.08	0.01	-0.04	0.03	0.04	-0.03	-0.01	0.35	0.79	0.45	<b>1.69</b>	-0.37
16	-0.14	0	-0.03	0.28	-0.19	-0.33	0.19	0.05	-0.26	0.36	0.37	-0.59	0.13	0.08	-0.01	-0.06	-0.21	0.27	0.73	0.01	-0.68
17	-0.25	0.09	-0.01	-0.19	0.06	0.25	0	-0.24	-0.24	-0.04	-0.04	0.15	0	0.03	-0.08	0.05	-0.52	0.9	0.79	<b>1.39</b>	-0.38
18	-0.79	0.03	-0.02	-0.11	0.01	0.11	-0.02	-0.09	-0.07	-0.44	-0.24	0.22	-0.1	0.07	-0.08	0.01	-0.43	0.45	-0.15	0.36	-0.46
19	0.12	0.04	0.01	-0.31	0.14	0.24	-0.14	0.52	-0.25	-0.45	-0.2	0.26	-0.07	0.07	-0.06	-0.01	-0.81	0.09	-0.46	0.26	-1.51
20	0.59	0.14	-0.08	-0.33	0.11	0.32	-0.24	-0.11	0.57	0.71	0.32	0.49	-0.14	0.07	-0.13	0.05	1.01	0.62	0.89	0.28	-0.74
21	-0.41	0.01	0.05	0.33	-0.11	-0.36	0.11	-0.39	-0.07	-0.25	-0.13	-0.01	0.07	0	0.02	-0.03	0.3	0.49	0.24	0.42	-0.59
22	0.12	0.08	-0.04	-0.18	0.08	0.32	-0.21	-1.41	0.96	0.22	0.1	0.18	-0.03	0.05	-0.08	0.03	-0.27	-0.03	0.42	-0.05	-1.2
23	-0.06	0.02	0.03	-0.04	0.01	-0.03	-0.02	0.09	-0.13	-0.26	-0.14	0.2	-0.01	0.07	-0.03	-0.04	0.66	0.7	0.63	<b>1.86</b>	-0.08

indicates that my algorithm is not only resulting in better probabilistic prediction, but also better point estimates compared to its base learners.

The multinomial logit model which predicts the probability of each base learner making the best prediction is built of 16 covariates (excluding the intercept) shown in Table 3.1. The hold-out-analysis resulted in training 72 various multinomial logit models. To understand their parameters and obtain some managerial insights, I derive an average over them through all the hold-outs. Table 3.1 summarizes these average values for the coefficients of this model. From Table 3.1, we can derive critical insights. Large positive variable coefficients is an indication for a direct relationship between that variable and the goodness of the corresponding model in prediction. For example, Table 3.1 indicates that model number 9 tends to make better predictions for windy days. Model number 12 preforms better than other models in the event of thunderstorms, and model number 1 outperforms other models in case of hail or tornadoes. The large positive coefficients for the two variables *mean snow amount* and *max snow amount* in the forth row indicate that model number 4 tends to make better predictions for winter storms. Similarly, the large positive coefficients for the mean QPF (Quantitative Precipitation Forecast) in the 11<sup>th</sup> row indicate that model number 11 makes better predictions for wet events. Having coefficients close to zero for the temperature-related variables indicates that these variables do not play a significant role in determining the best model. Finally, we see that model 2 performs better in cluster 2, model 3 and 10 perform better than others in cluster 3, and model 4 is the best model in cluster number 4. We also see that there are several models making good predictions in cluster number 5, but only a few of them perform well for cluster 6.

### 3.5 Summary and Conclusions

In this study, my goals were to develop a new BMA model for predictive modeling and address model uncertainty in the field of power outage prediction modeling. I developed a new two-stage adaptive algorithm based on Bayesian model averaging and used this algorithm for modeling daily customer interruptions. My approach had three main characteristics. (i) To implement BMA, I considered a decision-theoretic approach and modeled weights of the base learners with an online multinomial logit model. Weights of the base learners are dependent on the feature of the instances. (ii) In my model, unlike the classical BMA approach, the base learners are updated gradually as more data are observed. (iii) I extended my algorithm for the case when data are divided into multiple clusters. This helped my model be able to handle more complex datasets.

I validated my algorithm based on daily customer interruptions data. This was the first

application of BMA approach in the OPM literature. The results of holdout analysis showed that my algorithm results in more accurate probabilistic prediction than the base learners individually. I also found that as more data are observed, more accurate predictions are made by the proposed BMA model. Another important property of my algorithm was the strong inferences we can make. It helped us understand the conditions under which each of the base learners performs well. My case study was also the first single all-weather model developed in power outage predictive modeling literature. This significantly could help utilities better plan resource needs, and increasing the rate of restoration.

Although my work is motivated by power system application, my methodology and insights can be implemented in other predictive modeling problems dealing with high model uncertainty. It can especially be used in the problems for which not much initial information is available regarding the true model, or multiple models perform well formulating the process and we are looking for robust predictions. In general, my methodology can be used in various fields of application including biological and medical sciences (e.g., [5]), economics and social sciences (e.g., [27]), and other physical sciences and engineering applications (e.g., [55]).

## CHAPTER IV

# An Assessment of Drivers of Power System Damage During Severe Weather <sup>1</sup>

### 4.1 Introduction

#### 4.1.1 Research Motivation

Severe weather events have the potential to cause significant disruptions to the electric power grid. The resulting damages are, in some cases, very expensive and time-consuming to repair and they lead to substantial burdens on both utilities and customers [51]. Some examples of these events include tropical storms, flood, wind-storms, and heat waves. In January 1998, for instance, a major ice storm resulted in thousands of utility poles breaking and consequently the loss of power for more than 5 million people in Canada and northeastern U.S. In September 2003, Hurricane Isabel caused 1.8 million customers of Dominion Virginia Power to lose power, and also thousands of poles, spans of wires and transformers had to be replaced [67].

An important part of managing weather-induced power outages is being properly prepared for them, and this is tied in with broader goals of enhancing power system resilience. Modeling impacts of extreme weather events on the power system is a critical part of pre-storm resiliency practices because it directly influences the decisions made prior to, during, and after the event [41]. Accordingly, in the last two decades, a wide range of studies have been conducted in the outage prediction modeling (OPM) area. However, despite their widespread use, there are a number of limitations with current studies. In the next section, I briefly review these studies and summarize the contribution of my work in regard to these shortcomings.

---

<sup>1</sup>Submitted to Reliability Engineering and System Safety as Kabir, E., Guikema, S.D., and Quiring, S.M, McRoberts, B., An Assessment of Drivers of Power System Damage During Severe Weather

### 4.1.2 My Contributions to the Literature

Power outages, defined as a non-transitory activation of a protective device, are recorded by utility companies through an automated outage management system. A single outage may be associated with a widely varying degree of physical damage to the power system, and it could affect a small number of customers or hundreds of customers. The primary assets that experience damage in the power distribution network are overhead conductors and distribution lines, transformers, and utility poles that support conductors and transformers.

Previous works in the area of OPM have focused on predicting the number of outages (e.g., [37, 41, 42, 44, 67]), customers without power (e.g., [53]), or power outage duration (e.g., [83]). This is due to the ease of collecting outage data by utilities through an automated system. The results of these models are mainly used to inform the customer community about the size of outages and their length before the event and to inform utility decisions during an event. These predictions are also useful for the utilities to demonstrate to regulators that they can predict the extent of damage to the power distribution network. However, from a decision-making perspective, these predictions are not very useful for utilities to decide how many crews they need, what type of crews to request, and where to deploy them. Also, customer-focused OPM are not useful for making system reinforcement decisions at the asset level.

In this study, I first move beyond the previous OPM approaches and focus on damage data for different asset types including overhead (OH) and underground (UG) conductors, OH and UG transformers, and utility poles. I study the impacts of different meteorological variables on the failure probability of these utility components. Direct estimates of the effect of various meteorological factors on damage to the power system provide a much stronger basis on which utilities can make decisions about system reinforcement (hardening) as well as the level of emergency response materials (e.g., poles and line) to keep on hand before an extreme weather event.

In previous studies, the focus has been primarily on modeling and predicting outage-related variables in advance of a storm. The literature lacks an inferential study in which the effects of various factors on damage data are investigated. Thus, the second contribution of this article is to focus on studying the associations between meteorological variables and damaged power system assets using the Bayesian belief network (BBN) analysis. Having sound, long-term estimates for the impact of different factors on the power system provides utility companies with a basis on which to make more informed asset hardening decisions and to better explain the reasoning for their decisions to the regulators and the public. This study provides actionable strategies for the utilities to find vulnerable components of their system and to perform cost benefit analysis.

### 4.1.3 Chapter Organization

This chapter is structured as follows. Section 5.2 describes the literature. In Section 4.3, I define the input data and describe the BBN analysis in detail. Section 4.4 presents the BBN that I developed, and the effects of various weather events on the power system. Section 4.5 provides a list of insights from my analysis. Finally, the chapter closes with the summary and conclusion section.

## 4.2 Literature Review

This chapter is closely related to two main domains of research, namely power system damage modeling and Bayesian belief network.

### 4.2.1 Modeling Power System Damage

A number of studies have been conducted in the literature for modeling weather-induced damages to the power system. These studies can be divided into two primary approaches (1) fragility-based models, and (2) statistical learning models. In the fragility-based models, for each individual system component (e.g., a pole), a transfer function, which is called the fragility function, translates the key aspects of the weather hazard (e.g., gust wind speed) into the conditional probability of damage for that component. The damage probabilities are then used to simulate a number of replications of the damaged components, with each being converted into a set of customers without power through a power flow or network connectivity model.

Three examples of this approach are Winkler et al. [123], Han et al. [43] and Zhai et al. [134]. Winkler et al. [123] extended the fragility curve approach for power system poles impacted by hurricanes. They combined the fragility curves with topology-based simulation to use a connectivity model to characterize the impact of hurricanes upon power system reliability. Zhai et al. [134] developed a method to create a realistic synthetic network for a community and to then simulate realizations of damages and outages. Han et al. [43] also used a fragility-based approach to estimate the hurricane-related pole damage in the distribution system for a case study service area. They used the Bayesian methods for updating the results of structural reliability models with observed failure data.

Statistical learning models are trained using historic data about the performance of power systems during the previous weather events. The data usually includes the amount of damage in defined geographic areas together with information about the utility system, environmental and meteorological conditions. Statistical models learn the relationships between these

variables and the power system damages, and can predict the impact of future weather events. Guikema et al. [38] developed and compared the out of sample error of various statistical and machine learning models for predicting the number of damaged poles in the event of a hurricane. They found that non-parametric regression and data mining models may provide a better basis for accurate prediction of hurricane damage. In this study, they also emphasized that having accurate, geographically detailed damage data from multiple hurricanes is a strong basis for developing damage models. However, their damage data were not sufficient to develop a model with strong predictive accuracy.

The literature for estimating physical damage to the power system is very limited. In a few cases that address the physical damage to the power system, the models are usually developed for physical damage to the poles only. Furthermore, in all of these studies, the damage models were developed for hurricanes. Therefore, there is no article studying the effect of weather events such as windstorm, snowstorm, heatwaves, and rain events on different power system components. In this study, I address the above-mentioned gaps, and by using BBN I study the influence of various events on the failure probability of various components of the power system.

#### 4.2.2 Bayesian Belief Network

Bayesian Belief Networks (BBNs) were introduced by Wright [124] and further developed by Pearl [86] and Shafer and Pearl [101]. BBNs are a widely-used graphical model that provides a structured representation of the relations between random variables in an uncertain domain. A BBN consists of a qualitative part, which is a directed acyclic graph (DAG), and a quantitative part, which is a set of conditional probability tables. Each node in the DAG represents a random variable, while directed arcs between nodes represent dependencies or causal relationships between the variables. The BBN then represents the joint probability distribution over the set of random variables  $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$  denoted by  $P(\mathbf{X}) = \prod_{i=1}^n P(X_i|PA_i)$ , where  $PA_i$  stands for the parent set of  $X_i$ . A variable  $X_j$  is called a parent for  $X_i$  if there is a directed arc from  $X_j$  to the child variable  $X_i$  [21, 57].

The structure of a BBN also illustrates the conditional independence amongst the variables. D-separation is the criterion that summarizes the correspondence between conditional independence and a certain BBN structure. An undirected path  $p$  is said to be *d-separated* (blocked) by a set of nodes  $\mathbf{Z}$  if and only if (i) the path  $p$  contains a chain  $X_i \rightarrow X_j \rightarrow X_k$  or a fork  $X_i \leftarrow X_j \rightarrow X_k$  such that the middle node  $X_j$  is in  $\mathbf{Z}$ , or (ii) the path  $p$  contains a collider  $X_i \rightarrow X_j \leftarrow X_k$  such that the middle node  $X_j$  as well as any of its descendants does not exist in  $\mathbf{Z}$ . Then, for any three disjoint node sets  $\mathbf{X}$ ,  $\mathbf{Y}$ , and  $\mathbf{Z}$  in a BBN,  $\mathbf{X}$  is said to be *d-separated* from  $\mathbf{Y}$  by  $\mathbf{Z}$  if and only if  $\mathbf{Z}$  blocks every path from a node in  $\mathbf{X}$  to a node

in  $\mathbf{Y}$ .

Two main aspects of learning a BBN are (i) network structure learning and (ii) conditional probabilities learning. Once the network structure is realized, the conditional probabilities can be obtained using the data. In this step, the exact maximum likelihood estimates can be calculated by counting frequencies in the dataset. Learning the network structure is thus the crucial part [109].

#### 4.2.2.1 Network Structure Learning

A Bayesian network structure can be learned either from data, if available, or from experts, or a combination of both. The task of learning the network structure from data is computationally non-trivial due to the large size of the space of possible DAGs and it grows super-exponentially in the number of variables (nodes). Various learning methods are developed for this NP-hard problem. The structure learning algorithms can be classified into two groups: (i) scoring-based and (ii) constraint-based methods.

Score-based methods evaluate the quality of BBN structures using a scoring function and selects the one that has the best score. Therefore, score-based methods have two main elements: scoring functions and search strategies. Bayesian Dirichlet score [45], minimum description length (MDL) [62], Bayesian information criterion (BIC) [99], Akaike information criterion (AIC) [1], normalized maximum likelihood function [95], and the mutual information tests (MIT) score [10] are commonly used score functions in these methods. Two classes of search strategies are local search strategies (e.g., greedy hill climbing, Max-Min Hill Climbing [114], and stochastic search [80]) and optimal search strategies (e.g., search strategies based on Branch-and-Bound [20], Dynamic Programming [85], and Integer Linear Programming BN [50]).

Constraint-based methods operate in two independent phases: (i) constraint identification, and (ii) edge orientation. In the first phase, they use a series of conditional hypothesis tests to learn conditional independence relations among the variables in the model. Following these constraints, in the second phase, they build a (fully or partially) directed Bayesian network structure that best fits those independence relations. Classical (e.g.,  $\chi^2$ , and  $G^2$  statistics [125]), Bayesian (e.g., *BDeu* [18]) and information theoretic (mutual information [14]) tests are the most commonly used hypothesis tests in practice. The performance of these algorithms is critically determined by the accuracy of the adopted statistical tests. Thus, they may not work well when there are insufficient or noisy data.

#### 4.2.2.2 Belief Propagation

Belief propagation (BP) is a technique for making inference in Bayesian networks. In BP, we estimate the marginal posterior distribution of some unobserved variables in the system, conditional on observed variables. Then, we find the value of the unobserved variable that maximizes the posterior distribution. In BP, there are two types of message passing: (i) message from parents to children, which is called forward BP and (ii) message from children to parents, which is called backward propagation. In this study, I use both types of message passing using techniques that are explained below.

**Matching for confounding control:** In order to study the influence of an independent covariate on a target variable, we can compare distributions of the target variable for different values of the covariate under study. Then, hypothesis tests can be done to identify if there is a significant deviation between the distributions. In such circumstances, if observational data are used, the existence of confounding variables can cause bias in the hypothesis test results. A confounder is referred to as a variable that influences both the independent variable and the target variable without being an intermediate cause in the causal pathway between the independent variable and the target variable.

Matching is a technique from experimental design literature that attempts to mimic randomization [96]. Rather than pooling the entire sample for statistical analysis, matching creates pairs of instances that are similar in terms of the confounding variables, but have different values for the independent covariate under the study. The matched records are then used for hypothesis testing and the rest of the data are thrown away. We typically cannot match the confounders exactly especially when we have multiple confounders and thus, we need a metric of closeness. A commonly used closeness metric is the Mahalanobis distance which is calculated as:  $D(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X} - \mathbf{Y})^T S^{-1} (\mathbf{X} - \mathbf{Y})}$ . This is the square root of the sum of squared distances between each covariate scaled by the covariance matrix  $S$ .

**Most Relevant Explanation:** Most relevant explanation (MRE) is a method for finding multivariate explanations for a given set of evidence in a Bayesian network [133]. The main idea is to traverse a trans-dimensional space containing all the partial instantiations of the target variables, and find the one instantiation that maximizes a relevance measure. For a set of target variables  $\mathbf{X}$ , each observed (known) subset of  $\mathbf{X}$  is called a partial instantiation. Potentially, MRE can use any measure that provides common ground for comparing the partial instantiations of the target variables. Generalized Bayes factor (GBF) is a commonly used measure and has been shown to provide a plausible measure for representing the degree of evidential support. In this case, let  $\mathbf{X}$  denote a set of target variables, and the vector  $e$  be the evidence on the remaining variables in the Bayesian network. Maximum relevant

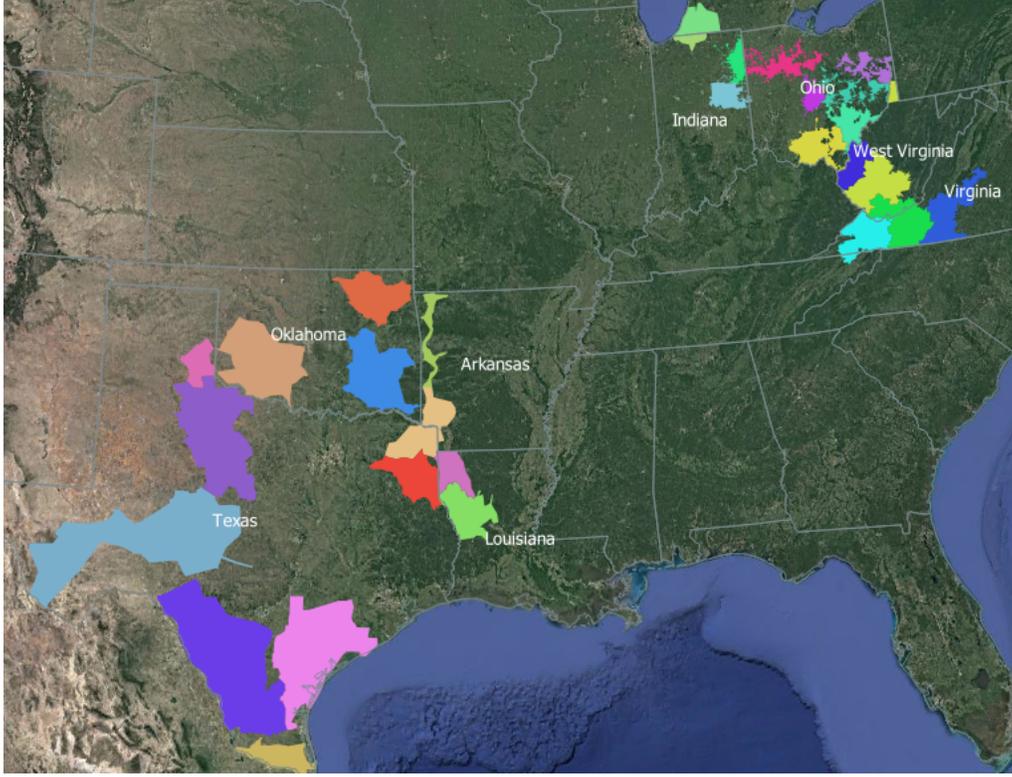


Figure 4.1: Spatial extent of the service territory for six operating companies serving 29 districts in 10 U.S. states. The total service territory is divided into 29 subareas that are defined by the AEP utility company and shown in different colors.

explanation is then defined as:

$$MRE(X, e) = \operatorname{argmax}_{x \subseteq \mathbf{X}, x \neq \emptyset} GBF(x; e),$$

where GBF is defined as  $GBF(x; e) = \frac{P(e|x)}{P(e|\bar{x})}$ .

**Mutual Information:** In information theory, the mutual information (MI) between two random variables quantifies the amount of information obtained about one random variable through observing the other random variable [16]. MI is a dimensionless quantity and can be thought of as the reduction in uncertainty about one random variable given knowledge of another. High MI indicates a large dependence between the two variables, while zero MI means the variables are independent. The MI of two random variables  $X_i$  and  $X_j$  is defined as:

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}.$$

Table 4.1: List of explanatory variables used for learning the model

Variable	Description	Source
Temperature (Mean of the eight 3-hour subperiods taken at each grid and averaged over the district area)	Forecast air temperature taken at a height of 2 meters above the surface.	Global Forecast System (GFS) model.
12-month standard precipitation index (daily value taken at each grid and averaged over the district area)	For a 12-month period, the number of standard deviations of aggregated precipitation amount above or below the median from a probability distribution function that is computed from historical data.	North American Land Data Assimilation v.2 (NLDAS-2)
Soil moisture (at the 40-100 cm layer)	At the given layer, the mean instantaneous volumetric water content taken daily and transformed to a percentile based on the cumulative distribution function of historical data taken from a 31-day window surrounding the given date.	NLDAS-2
Snow depth (Mean of the eight 3-hour subperiods taken at each grid and averaged over the district area)	The depth of snow on the surface (m)	GFS model.
Total rain (24-hour total taken at each grid and averaged over the district area)	Total amount of precipitation accumulated (kg/m <sup>2</sup> ) in the form of rain.	GFS model.
Wind gust speed (Mean of the eight 3-hour subperiods taken at each grid and averaged over the district area)	The maximum instantaneous wind speed (m/sec) forecast; it does not account for localized wind gusts resulting from thunderstorms.	GFS model
Surface lifted index	The forecast difference between the observed temperature at 500 hPa and the temperature of an air parcel lifted to 500 hPa from near the surface. Negative values indicate an unstable environment, with instability increasing as the magnitude negative values increase.	GFS model
Convective precipitation (24-hour total at each grid and averaged over the district area)	Total amount of liquid precipitation (kg/m <sup>2</sup> ), caused by unstable air (commonly in the form of thunderstorms).	GFS model
Sea level pressure (Mean of the eight 3-hour subperiods taken at each grid and averaged over the district area)	The atmospheric pressure reduced to mean sea level (Pa)	GFS model
Dewpoint (k) (Mean of the eight 3-hour subperiods over the total subarea)	A forecast of the temperature (K) to which air must be cooled to become saturated with water vapor.	GFS model
General thunder	A forecast representing the chance of a thunderstorm.	National Weather Service Storm Prediction Center
Storm relative helicity	The cyclonic updraft rotation in right-moving supercells (m <sup>2</sup> /s <sup>2</sup> ), calculated for the lowest 1-km and 3-km layers above ground level.	GFS model
Convective inhibition (Mean of the eight 3-hour subperiods over the total subarea)	The amount of energy (J/kg) needed to lift an air parcel from the lifting condensation level to the level of free convection.	GFS model
Absolute vorticity - 500 mb (Mean of the eight 3-hour subperiods taken at each grid and averaged over the district area)	The strength of rotation in the atmosphere (sec-1). Positive values represent cyclonic rotation and negative values anti-cyclonic rotation.	GFS model
Severe thunderstorm risk (daily value taken at each grid and averaged over the district area)	A binary forecast where a value of one indicates that there is an enhanced risk of severe thunderstorms; Qualitatively as "isolated severe thunderstorms possible" and quantitatively as "5% probability of severe storms within a 25-mile radius".	National Weather Service Storm Prediction Center

### 4.3 Data Description

The daily damage data used for training the Bayesian network is acquired from six operating companies from 2016 to 2020. They serve 29 utility-defined subareas in Michigan, Indiana, Ohio, Kentucky, West Virginia, Virginia, Tennessee, Arkansas, Louisiana, Texas, and Oklahoma. Figure 4.1 shows the spatial extent of their service territory. The subareas are defined by the utility companies serving the area and I model at these subarea levels.

The damage data that forms the basis for my model consists of the daily percentage of damaged poles, UG and OH transformers as well as the number of damage in OH and UG conductor divided by the number of miles of OH and UG conductors, respectively. These damage proportions are recorded for each subarea. The proportion of daily customer interruptions for each subarea is also recorded. The utility provides monthly refreshes of the damage and customers interrupted data, with each district typically recording several hundred events leading to a loss of transmission to customers each month. For every single event, the initial outage time, the type of equipment damaged, the number of customers affected, and duration of the event are recorded. These data are aggregated spatially to the district-level and assigned to a single calendar day based on the start times of the outage events. The damage and customer interruption data are combined with other explanatory variables described in Table 4.1.

To develop a BBN using the above-described data, I transform each continuous variable into a categorical one with at most five categories (levels). Table 4.2 shows the range of values in each category. To divide each variable into the five levels, the breaking points are selected such that the number of records is approximately the same in all the five categories.

**Variable Selection:** The 15 explanatory variables were selected, because they were deemed the most influential on the response variables. Here, being influential means that a change in one of these 15 variables will have a more significant effect on the damage variables than any of the other variables considered. To test for influence, I consider two factors including each variable's impact on: (i) increasing the overall BIC score of the network, and (ii) reducing the uncertainty about the damage variables. After repeating the forward adding and backward elimination process based on variable importance, I observe that these 15 variables appeared with a high level of importance for both the network and the damage variables.

Table 4.2: List of all variables with their ranges. Each continuous variable is transformed into a categorical one with at most five levels. Each column of the table indicates the range of values falling in the corresponding category (level).

	Level 1	Level 2	Level 3	Level 4	Level 5
Damaged OH conductors	(0, 0]	(0, 2.1e - 4]	(2.1e - 4, 4.0e - 4]	(4.0e - 4, 7.1e - 4]	(7.1e - 4, 1.3e - 1]
Damaged UG conductors	(0, 0]	(0, 9.3e - 4]	(9.3e - 4, 1.7e - 3]	(1.7e - 3, 3.2e - 3]	(3.2e - 3, 1.5]
Damaged poles	(0, 0]	(0, 4.4e - 6]	(4.4e - 6, 5.8e - 6]	(5.8e - 6, 1.0e - 5]	(1.0e - 5, 6.7e - 4]
Damaged OH transformers	(0, 0]	(0, 1.8e - 5]	(1.8e - 5, 2.7e - 5]	(2.7e - 5, 3.7e - 5]	(3.7e - 5, 1.4e - 3]
Damaged UG transformers	(0, 0]	(0, 4.4e - 5]	(4.4e - 5, 7.9e - 5]	(7.9e - 5, 1.4e - 4]	(1.4e - 4, 3.7e - 2]
Customer interruptions	(0, 1.9e - 4]	(1.9e - 4, 6.0e - 4]	(6.0e - 4, 1.6e - 3]	(1.6e - 3, 5.7e - 3]	(5.7e - 3, 8.2e - 1]
Temperature	(2.53e2, 2.78e2]	(2.78e2, 2.85e2]	(2.85e2, 2.92e2]	(2.92e2, 2.97e2]	(2.97e2, 3.10e2]
12-month SPI	(-2.5, -2.6e - 1]	(-2.6e - 1, 7.6e - 2]	(7.6e - 2, 3.8e - 1]	(3.8e - 1, 7.6e - 1]	(7.6e - 1, 2.7]
Soil moisture	(0, 2.9e - 1]	(2.9e - 1, 4.6e - 1]	(4.6e - 1, 6.3e - 1]	(6.3e - 1, 8.0e - 1]	(8.0e - 1, 1]
Snow depth	(0, 0]	(0, 3.6e - 1]	NA	NA	NA
Total rain	(0, 0]	(0, 1.5e - 2]	(1.5e - 2, 4.9e - 1]	(4.9e - 1, 3.5]	(3.5, 1.22e2]
Wind gust speed	(9.4e - 1, 3.5]	(3.5, 5.2]	(5.2, 7.0]	(7.0, 9.2]	(9.2, 2.3e1]
Surface lifted index	(-1.2e1, -1.4]	(-1.4, 3.4]	(3.4, 1.0e1]	(1.0e1, 1.7e1]	(1.7e1, 3.9e1]
Convective precipitation	(0, 0]	(0, 3.4e - 2]	(3.4e - 2, 4.1e - 1]	(4.1e - 1, 2.8]	(2.8, 8.2e1]
Sea level pressure	(9.924e4, 1.012e5]	(1.012e5, 1.015e5]	(1.015e5, 1.018e5]	(1.018e5, 1.022e5]	(1.022e5, 1.046e5]
Dewpoint	(2.48e2, 2.73e2]	(2.73e2, 2.80e2]	(2.80e2, 2.87e2]	(2.87e2, 2.92e2]	(2.92e2, 2.99e2]
General thunder	(0, 0]	(0, 1.4e - 1]	(0, 1]	NA	NA
Storm relative helicity	(-1.1e2, 4.7e1]	(4.7e1, 8.0e1]	(8.0e1, 1.2e2]	(1.2e2, 1.8e2]	(1.8e2, 9.1e2]
Convective inhibition	(-7.0e2, -4.6e1]	(-4.6e1, -1.3e1]	(-1.3e1, -8.2e - 1]	(-8.2e - 1, 1.9]	(1.9, 5.77e1]
Absolute vorticity-500mb	(-2.6e - 5, 5.9e - 2]	(5.9e - 2, 7.2e - 5]	(7.2e - 5, 8.6e - 5]	(8.6e - 5, 1.1e - 4]	(1.1e - 4, 3.8e - 4]
Severe thunderstorm risk	(0, 0]	(0, 1]	NA	NA	NA

## 4.4 Method and Computational Results

### 4.4.1 Network structure learning

The aim of this research is to develop a Bayesian network for assessing and quantifying the resilience of an electric power system. I use daily damage data for 29 districts in 10 states shown in Figure 4.1, from 2016 to 2020. I use daily damage data for poles, OH and UG transformers, and OH and UG conductors because it provides us with the opportunity to assess the impact of different weather events on the power system. Different parts of a power network may experience disruptions due to weather events. Damage in any of these assets can result in loss of power for a number of customers. I also study the daily number of customers without power (customer interruptions).

Figure 4.2 demonstrates the Bayesian belief network that I develop to represent the associations between meteorological variables and the damaged assets and customer interruptions in the power grid. To learn this Bayesian network structure and the conditional probability functions, I respectively use the *hc()* and *bn.fit()* functions from the *bnlearn* package in R [100]. The *hc()* function learns the network structure based on the score-based hill-climbing algorithm. The hill-climbing algorithm usually starts from an empty network without any edge, or a randomly generated structure, and then iteratively applies single edge operations, including addition, deletion and reversal, looking for the choice that locally maximizes the score improvement. I selected BIC as the score function for this algorithm. The *bnlearn*



Figure 4.2: The Bayesian belief network representing the associations among meteorological variables (shown by white circles), damaged assets (shown by yellow circles) and customer interruptions (shown by brown circle) in the power grid.

Table 4.3: List of the features I study their effects on the damage and customers interrupted variables using forward BP. For each studied feature, I also show its confounders that have backdoor path to the damage variables (i.e., variables that influence both the studied feature and the damage or customers interrupted variables without being an intermediate cause in the causal pathway between the studies feature and the damage or customers interrupted variables). The confounders are discovered using the BBN shown in Figure 4.2.

Studied Features	Confounders
Total rain	Temperature, Surface lifted index
Snow depth	Temperature, Wind gust speed
Temperature	12-month SPI
12-month SPI	-
Soil moisture	Temperature, 12-month SPI, Total rain
Wind gust speed	Temperature
Surface lifted index	Temperature

package allows us to define black and white lists indicating the arcs that should be avoided by the algorithm as well as the arcs that should certainly be included in the network. This property of the *bnlearn* package and the *hc()* function allows us to combine the expert knowledge with data-driven algorithm in order to learn the network structure. In this regard, I first force the network to avoid any arcs from the damage and customer interruption variables to the weather forecast nodes. Second, I force the network to include arcs from each damage variable to the customer interruption node. The latter is because I believe that each damage can cause a number of customers losing their power. The rest of the arcs in the network shown in Figure 4.2 are data driven using the hill-climbing algorithm.

## 4.4.2 Belief Propagation Analysis

### 4.4.2.1 Forward propagation

In performing inference on the network, my goal is to study the effects of various weather conditions on the failure of power system. The weather events that I study include (1) high rain (level 5), (2) high snow (level 2), (3) low (level 1) and high (level 5) temperature, and (4) low (level 1) and high (level 5) 12-month precipitation index, (5) low (level 1) and high (level 5) soil moisture, (6) high wind speed (level 5), and (7) low surface lifted index (level 1). The first column of Table 4.3 indicates the variables that describe these different weather conditions. I study and compare the effects of these variables on the failure of different parts of the system as well as the customer interrupted variable.

Before performing the analysis, I need to investigate if there exists any confounding variable that may cause bias in the results. The confounding variables that I need to control



Figure 4.3: Results of the ANOVA test for each variable and damage type. The boxplots show the distribution of damages in various levels of each variable. The connected red points indicate the mean damage in each group. P-value of the ANOVA test, the maximum and minimum damage means are also shown in each sub-figure.

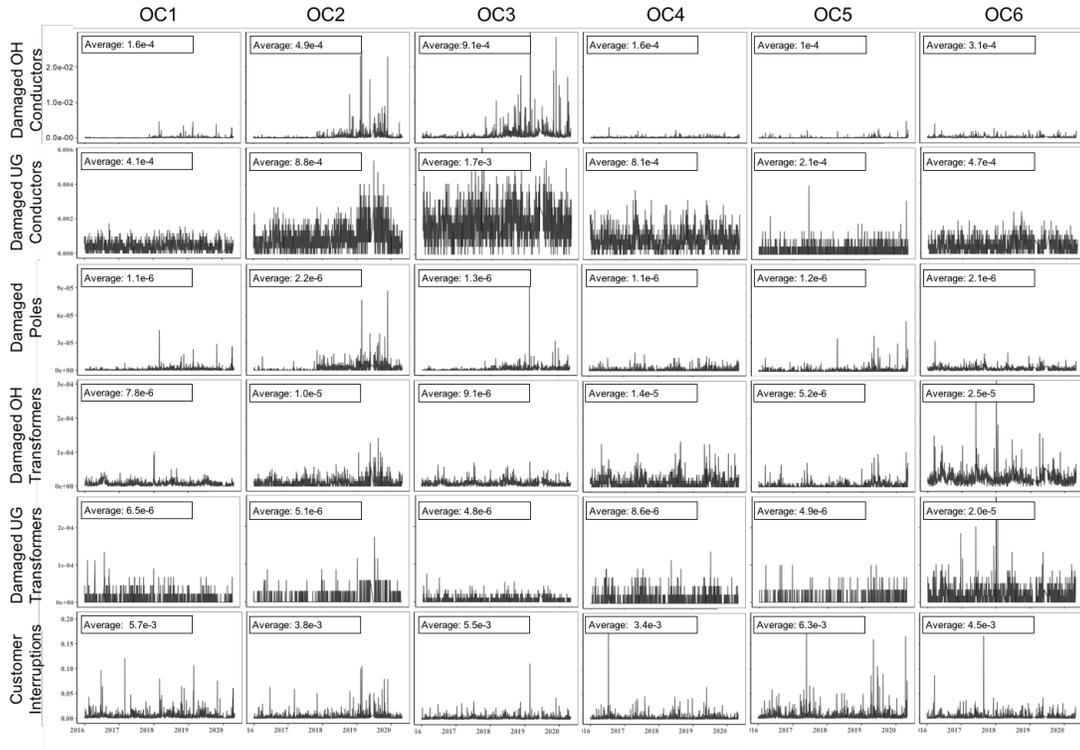


Figure 4.4: The daily proportion of different damage types in the service territory of each operating company.

for are those predecessors, which have backdoor path to the damaged variables. I find these confounders given the Bayesian network shown in Figure 4.2. The second column of Table 4.3 shows the confounders that I should control for. For a given variable and damage type, after I apply matching for confounders, I conduct a one-way analysis of variance (ANOVA) to see if there exists a statistically significant difference between means of damages given various levels of that variable. The result of the ANOVA test for each variable and each damage type is shown in Figure 4.3. In each test, the null hypothesis is that the means of damages are the same in different levels of that variable, and the alternative hypothesis is that the damage mean in at least one of the levels of that variable is not equal to the damage means of other levels. In Figure 4.3, each sub-figure represents the result of each hypothesis test as well as the boxplot of damage sizes in different levels of the variable. The p-value, maximum and minimum means are also reported for each test.

From Figure 4.3, we observe that higher rain results in more damage to all components of the power system, and consequently, more customer interruptions. This increase may be due to an increase in soil moisture, which reduces its stability, or due to extra pressure/force on trees and poles that can cause them breaking (that can also cause damage to OH conductors and transformers). We also find that a high level of damage to the UG and OH

transformers occurs when rain level is equal to its minimum level. High snow (level 2 snow depth) on the other side does not increase chance of failure in OH conductors, OH transformers and poles. However, it results in more damage to UG conductors and UG transformers, and consequently, more customer interruptions. Studying the temperature variable indicates that OH and UG transformers are very vulnerable to temperature. Increase in temperature significantly enhances damage rate of these two components. It also makes some enhancement in the damage rate of OH and UG conductors and finally, results in higher customer interruptions. My results show that low temperature is also associated with higher damage level in OH and UG conductors. We know that low temperature is associated with high rain and soil moisture and these two increase damage rate of OH and UG conductors. Thus, low temperature, indirectly, increases damage rate in these assets.

12-month SPI represents long term (one year) precipitation condition. Figure 4.3 shows that higher precipitation during 12-month period significantly increases the chance of damage in OH and UG conductors and poles, but surprisingly, it does not result in higher customer interruptions. My results also show that damage rate of UG conductor is also high in the days with lowest 12-month SPI. This may be because of the impact of 12-month SPI on the temperature variable. Higher soil moisture enhances damage rate of UG conductors and UG transformers. Very low (level 1) and very high (level 5) soil moisture are also associated with higher pole damage and they end up with more customer interruptions. Higher wind gust speed increases damage level of OH conductors, poles and OH transformers. This increase is more significant in OH conductors than other parts of the system. Higher wind gust speed also causes a higher rate of customer interruptions. The surface lifted index variable has a significant impact on the OH conductors and poles. The lower level of this variable indicates highly unstable weather condition and is associated with significantly more damages to these two assets. This leads to a significantly higher rate of customer interruptions. We also see that the influence of this variable is much more than the wind gust speed variable, especially in pole failures.

Figures 4.4 and 4.5, respectively, represent the daily proportion of different damage types and the distribution of various weather features in the service territory of each operating company over time. In these figures, I remove the names of the operating companies for the purpose of protecting identity of the individual operating companies and represent them with the names: OC1, OC2, ..., OC6. Studying and comparing Figures 4.3, 4.4 and 4.5 help us better detect the vulnerable components as well as the main damage causing factors in the power system for each operating company.

Figure 4.4 shows that in OC1, damage rate in all asset types as well as customer interruptions are in the middle level compared to other OCs. OC2 has the second rank of

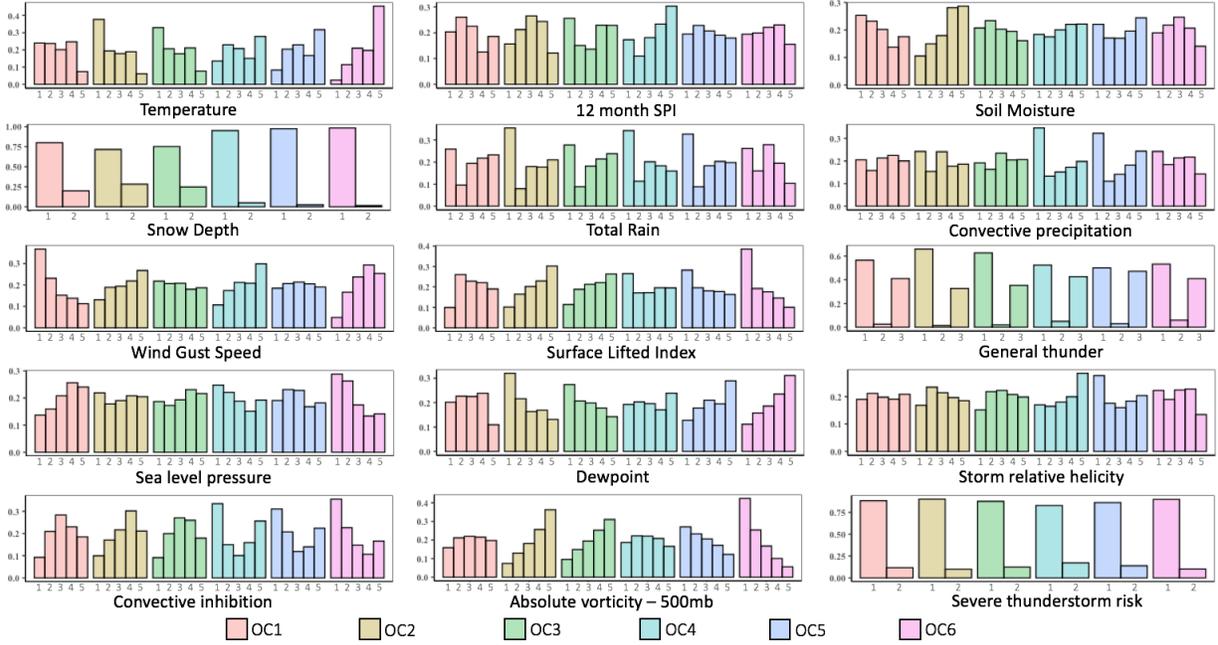


Figure 4.5: The distributions of different variables in the service territory of each operating company.

damaged OH and UG conductors and first rank of damaged poles. High rain, snow depth, soil moisture and wind in the service territory of this operating company are the potential reasons for high rate of damaged OH and UG conductors, and poles. Although this operating company has high rate of damage in these asset types, it has the second lowest rate of customer interruptions among all operating companies. This shows that their power system is more resilient than others.

OC3 has the highest rate of damaged OH and UG conductors and second rank of customer interruptions among all OCs. The service territory of this OC experiences high rain, high snow and low temperature compared to other OCs and these factors may be the potential reasons for the high damage rates. My results also show that although OC2 experiences higher winds and more unstable weather, its damage rate of OH and UG conductors is less than OC3. OC4 has the second rank in damage rate of OH and UG transformers. High temperature and low rain are the potential reasons for these damages. This OC also experiences highest wind speed and low surface lifted index, but rate of damage in OH conductors and poles is not significant compared to others. Studying damage rate of different components in OC5 shows that even though ratio of damaged assets in this OC is not high compared to other OCs, it experiences the highest rate of customer interruptions. This may be due to not recording damage data precisely or having a very vulnerable power system in which a damage results in loss of power for a large number of customers. OC6 experiences

Table 4.4: Summary of the backward belief propagation results. In each analysis, I set a damage/outage variable equal to its highest level and compute the generalized Bayes factor (GBF) for every combination of the set of target variables. I then report the condition under which the GBF value is maximized. In some cases, more than one condition may result in maximum GBF.

Evidence	MRE							Weather Description	Max GBF
	Total rain	Snow depth	Temperature	12-month SPI	Soil moisture	Wind gust speed	Surface lifted index		
OH conductors = 5	4	2	2	3	4	5		snowy, windy rainy, windy, unstable weather	10.1
	5		3	3	2	5	1		
UG conductors = 5		2		5	5	1	1	unstable weather, saturated soil	15.5
Poles = 5			3	2	5	5	5	windy, saturated soil snowy, windy	25.3
		2	1	4	2	4	3		
OH transformers = 5			4	5	2	4	4	warm, windy snowy, windy	14.5
	4	2	2	4	3	5	4		
UG transformers = 5		1	1	2	5	3	3	cold, saturated soil warm, rainy	72.5
	5	1	4	4	3	2	2		
Customers interrupted = 5		2	2	3	3	4	4	hot, saturated soil snowy, windy cold, rainy	5.3
	5	1	1	3		2	5		

the highest rate of damages in OH and UG transformers. In this area, frequency of very high temperature and low rain is higher than other places and these two factors may be the reasons for these damages. Although high wind speeds and unstable weather conditions are very frequent in this area, OC6 does not have significant rate of damaged OH conductors and poles. Thus, we can conclude that high wind and unstable weather are not by themselves the deriving causes of damages in poles and OH conductors. But, they are dangerous when combined with high rain, saturated soil and cold weather as we see in OC2 and OC3.

#### 4.4.2.2 Backward propagation

In the backward belief propagation analysis, I set each damage/outage variable equal to its highest level, and find the MRE. Table 4.4 summarizes the results. Note that the minimum possible value of GBF is one and it happens when the evidence and the values of the target variables are independent from each other. The higher the GBF is, the more association exists between the values of the target variables and the evidence. Unlike hypothesis test method in which I study the effect of each individual variable on the failure of the power system, in this approach, I find the weather events that have the highest association with the highest rate of damage to the power system. Each weather event is represented by multiple weather factors. Table 4.4 shows that the highest rate of damaged OH conductors is observed under (i) warm and rainy event and (ii) snowy and windy events. The highest rate of UG conductor damages is seen under snowy and unstable weather with saturated soils. Pole damages are also found more often under (i) windy and saturated soil and (ii) snowy and windy conditions. The highest rate of OH transformer damages is observed under (i) warm

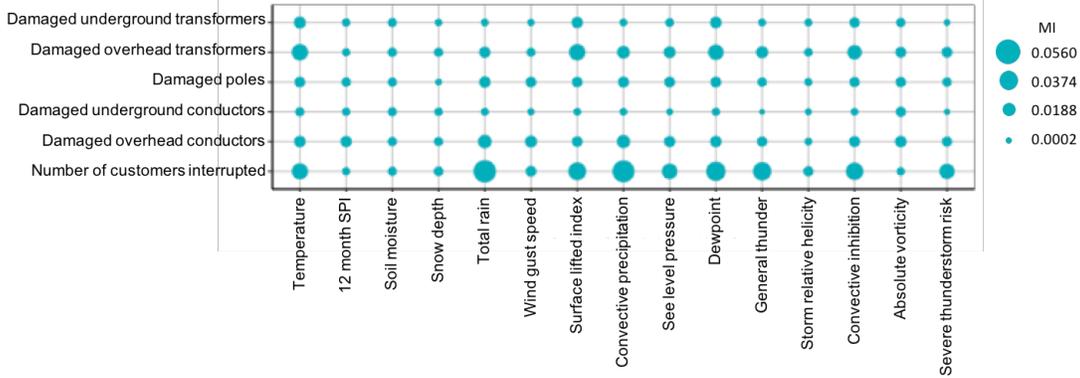


Figure 4.6: The comparison of the MI between different meteorological variables and damage/customer interruptions variables. The larger the size of the circle is, the larger the MI exists between the two corresponding variables.

and windy, and (ii) cold and windy events. The highest damage rate of UG transformers is seen under both cold events with saturated soil and rain events with warm weather condition. Finally, since each instance of damage results in loss of power for customers, we see that different weather conditions including (i) hot weather with saturated soil conditions, (ii) snowy and windy, and (iii) cold and rainy are the most common explanations for the highest level of customer interruptions.

#### 4.4.2.3 Variable Importance

To find the importance and influence of different variables on the power system damage and customer interruptions, I estimate the mutual information (MI) between each variable and the damage or customer interruption variable. Figure 4.6 compares the size of the MI between different variables and the damage/customer interruption variables. The larger the size of each circle is, the larger the MI exists between the two corresponding variables. Table 4.5 for each damage/outage type, ranks the variables based on their MI. Lower numbers represent higher MI between the two corresponding variables. Figure 4.6 and Table 4.5 show that temperature, surface lifted index and convective inhibition have the highest MI with OH and UG transformers. Total rain and convective precipitation are the two variables having highest MI with both damage OH conductors and poles. Absolute vorticity has the maximum MI with damaged UG conductor and finally, total rain and convective precipitation are the most important features for the rate of customers interrupted.

Figure 4.6 shows that damaged UG conductor has the lowest level of MI with the variables, while customers interrupted variable has the highest level of MI. Table 4.5 shows that storm relative helicity, snow depth, and severe thunderstorm risk have respectively the

Table 4.5: Ranking of the variables based on their MI with the damage data. Lower numbers represents higher MI between the two corresponding variables.

Damage/outage type	Temperature	12 month SPI	Soil moisture	Snow depth	Total rain	Wind gust speed	Surface lifted index	Convective precipitation	Sea level pressure	Dewpoint	General thunder	Storm relative helicity	Convective inhibition	Absolute vorticity	Severe thunderstorm risk
Damaged underground transformers	1	8	6	13	12	14	3	11	7	2	10	9	4	5	15
Damaged overhead transformers	1	15	11	12	8	13	2	5	7	3	6	14	4	10	9
Damaged poles	5	11	13	15	1	6	7	2	4	3	12	14	9	8	10
Damaged underground conductors	2	5	3	4	7	11	8	9	13	6	15	12	10	1	14
Damaged overhead conductors	6	7	13	14	1	3	4	2	8	10	12	15	9	5	11
Customer interruptions	7	15	13	12	1	5	3	2	8	10	4	11	6	14	9

lowest mutual information with the damage/outage variables and thus, are the least important variables in the model. On the other side, temperature, total rain, surface lifted index and convective precipitation have the highest level of mutual information with the damage/outage variables.

## 4.5 Discussion and Insights

Severe weather events lead to damage to power system components and result in customer interruptions. In this chapter, I studied the impacts of various weather conditions on five different components of the power system: OH and UG conductors, poles, OH and UG transformers in the service territory of six operating companies serving 29 districts within 10 U.S. states.

In this chapter, I provided a list of insights, which can help utility companies understand the factors driving outages, find the vulnerable components of power systems and suggest actionable strategies for the utilities to perform cost-benefit analysis. The first insight was that the areas that experience high rain, snow and wind have higher rate of damage to OH conductors and poles. In such areas, utilities may decide to perform hardening practices such as vegetation management, replacing wooden poles with poles that can withstand higher wind speeds, or replacing OH conductors with UG conductors if there is not high risk of flooding in the area. We found that high wind and unstable weather are not by themselves the deriving causes of damages in poles and OH conductors. However, they are dangerous when combined with high rain, saturated soil and cold weather. The second insight is that extreme heat especially when paired with rising humidity levels are very dangerous and cause more damage to the OH transformers. This implies that in the areas with such

weather conditions, utilities should consider upgrading their transformers. The next insight is that UG conductors are vulnerable to very cold, rainy, and saturated soil conditions and thus, in the areas that experience these conditions more often, undergrounding conductors are better to be avoided.

There are many challenges in gathering power system damage data, and that is why not many studies have been done with a focus on weather-induced power system damages [38]. In this study, I used a dataset of daily damage to the power system covering 29 districts in 10 U.S. states. However, there were still a few limitations with my dataset. One issue was the fact that the recorded damages could be caused by any factor (e.g., car accident, or animals) and not just meteorological factors. Even though the influence and frequency of the non-weather-related damages were not substantial, it might cause some errors in the developed model and results. The other issue was not including the energy consumption data in the model. Under severe weather events, like very hot or very cold conditions, customers might consume more energy and the increase in energy consumption might cause damage to the power system and result in customer interruptions. In such situation, the weather by itself might not be the reason for damages to the system. Therefore, in future studies, including energy consumption in the model may help professionals understand the effect of various factors more precisely.

I suggested that operating companies should pay more attention on collecting the damage data more precisely in their service territory. This data can help better evaluate the resiliency level of their system and find the vulnerable component that need more focus. If in an OC a few damages result in a large proportion of customers losing their power, the OC should invest more on improving the resilience of their power network. I also suggested that deploying resources such as back-up distributed generators, automatic tie switches, physically controlling power flow in distribution networks, and self-healing schemes are the effective resiliency strategies that can be helpful in these situations.

## 4.6 Summary and Conclusions

Extreme weather events, from winter storms to heat waves, impact the power system and are potential to cause significant damages. Current climate models indicate that the risk from extreme weather events is severe and has increased in the recent years. Frequency and intensity of hazards such as high winds, heavy precipitation, and prolonged heat events have also increased over the past years. Currently, extreme weather events are the main cause of damage to the power system and consequently electric power outages in the U.S. To mitigate these risks, utility companies invest millions of dollars every year for hardening

the power systems and improving their resilience.

In this study, I developed a Bayesian belief network, which represents the interconnection between various meteorological factors and damages to different power system components. To estimate the impacts of the meteorological variables, I conducted hypothesis tests. To control the effects of confounding variables on the hypothesis test results and mimic randomization in the data, I applied a matching technique before performing the hypothesis tests. This approach helped us to understand the effects of each individual variable on the power system damages. In addition to the forward belief propagation (i.e., hypothesis tests), I conducted the backward belief propagation using the maximum relevant explanation technique. In this method, I investigated the weather conditions that derived the maximum level of damages to the different parts of the power system. Unlike the first approach, it showed the combined weather conditions that causes maximum level of damages. Finally, I performed variable importance analysis to rank the meteorological factors based on their influence on the power system damages.

This study was based on a real dataset of daily damages occurring in 29 districts of 10 U.S. states from 2016 to 2020. These districts were served by six operating companies. The results of my analysis found that temperature, total rain, surface lifted index and convective precipitation are the most important variables identifying the level of damage to the power system. It also suggested that the UG conductors are more susceptible to cold weather conditions with high soil moisture, while damaged UG transformers are caused under both warm and cold events with high soil moisture. I also found that high wind by itself does not cause significant damage to the system, but when it is combined with high rain, snow and soil moisture, it becomes very dangerous.

Among the studied operating companies, OC2 looks more resilient than others. Although it experiences the worst weather and highest damage rates in its power system components, it has the second lowest rank of customer interruptions. On the other side, OC5 seems to be the least resilient. Even though weather conditions in this area are milder than other studied places and the operating company does not experience large number of damages, it has the highest rate of customer interruptions. Thus, my results suggested that this company should invest more on improving the resilience of their system by proper practices.

## CHAPTER V

# A Multi-stage Stochastic Crew Coordination Model for Power Outage Restoration <sup>1</sup>

### 5.1 Introduction

Natural hazards often cause significant economic and physical disruptions to energy infrastructures and lead to substantial inconvenience for residents living in the impacted areas due to loss of electricity [123]. The growing number of people affected by natural hazards, the inherent uncertainty and complexity of such phenomena and the difficulties they cause, establish the necessity for utility companies to make better measures and practices in order to reduce risk and environmental damage of these events on the power system [53]. However, this is not an easy task considering the large uncertainty these phenomena present. As reported by the International Disaster Database, the total number of natural disasters appears to be growing, as well as the number of people affected by them. This may be influenced by problems ranging from limited resources and delay in the arrival of these resources (e.g., maintenance crews and materials), huge uncertainty in response times, lack of emergency planning, and demand uncertainty.

Some studies have been presented in the context of emergency planning for power systems in the face of natural hazards. Almost all of these studies focus on developing a scheduling model for maintenance crews and sequencing and routing to disrupted network components in order to optimize restoration process given the knowledge about the power network and available resources. However, in the case of extreme events, such as hurricanes, local resources are not sufficient for restoring power in a reasonable time and utilities have to request extra resources, in particular repair crews, from other utilities. The resource allocation decisions should be initiated prior to the event because in most cases it takes time until external crews arrive to the impacted areas. These decisions significantly affect other resiliency decisions

---

<sup>1</sup>We intend to submit a modified version of this chapter as a journal paper.

made by the utility companies prior, during and after the event (e.g., scheduling, sequencing and routing decisions). To the best of my knowledge, there are few optimization models (e.g., [117]) in the literature concentrating on these types of resource allocation decisions.

In addition to the above mentioned shortcoming, a number of research gaps have been highlighted by recent survey papers [47, 35, 32] and my literature review on optimization methods developed in disaster operations management. First, most optimization models are single-objective. However, there exist many conflicting objectives such as minimization of total costs on the one hand and maximization of satisfied demand on the other, that are important to be optimized simultaneously as a multi-objective optimization model. The use of multi-objective models for decision making is appropriate considering the different actors involved in the decision process. Second, although many papers in this area design a one-echelon network, such models reflect reality in a limited way. Because in general the network contains more than one participant, more realistic models can be achieved by considering multi-echelon network including, for example, headquarter companies, utility companies and staging areas. Such an extension to a multi-echelon model is necessary for establishing a more realistic counterpart to the single-echelon model. A multi-echelon network can help authorities serve the affected people more properly and in a timely manner. Third, in the existing literature, the possibility of inter-facility crew and material transfers is not considered. Each distribution center has to serve a specific number of demand areas, but the possibility of transporting some resources to another distribution center in case of immediate needs has not been considered yet this is common in practice. forth, most optimization models consider a single-period framework and use the two-stage stochastic programming (TSSP) approach as a common way of dealing with uncertainties for their single-period framework. However, the use of a multi-period model helps the decision maker because it is a more comprehensive analysis and new information (once it is known) can be included for the future periods. Further, considering multi-period models, the common TSSP framework can also be expanded to multiple stages, where additional information can be incorporated in the model formulation allowing a more detailed decision process.

To address the above concerns, I propose a comprehensive and generic optimization approach for establishing staging areas, allocation and relocation of repair crews between utility companies, staging areas, utility stores, and other contracting agencies. This approach includes a new multi-stage and multi-echelon stochastic programming model. My optimization model determines which staging areas should be opened in advance of a disaster, and how supplies (including repair crews with vehicles provided by various utility companies) should be pre-positioned in staging areas and distributed among a network of demand centers and other distribution centers. I also consider lead times between crew transshipments used for

sharing. I consider and evaluate two conflicting objectives including (1) minimizing total restoration cost, and (2) maximizing the expected utility levels of demand points concurrently (unlike all single objective models in the literature). My model can mitigate each of the issues mentioned above through a new multi-stage stochastic programming (MSSP) model.

The rest of this chapter is organized as follows. Section 5.2 summarize the literature on optimization models for improving the restoration process in power systems. Section 5.3 explains the problem and Section 5.4 describes the proposed mathematical model for my crew coordination problem. Section 5.5 illustrates the numerical results and performance of the proposed model. Finally, Section 5.6 provides the concluding remarks.

## 5.2 Literature Review

Several studies in recent years have focused on developing optimization models and algorithms to improve the restoration process of power systems after disruptive events. Many of the fundamental studies in the field of post disruption power infrastructure resilience focus on scheduling and sequencing disrupted network components to restoration crews. Kim et al. [60] developed a mixed integer programming model to minimize the weighted sum of total damage while considering a repair crew problem in which aspects of damage vary at certain times. Nurre et al. [103] introduced an integrated network design and scheduling problem to improve the infrastructure network construction and restoration process. They developed a heuristic dispatching rule to identify the next set of network components to be restored by crews in order to maximize the cumulative weighted flow in the network over a horizon. Sharkey et al. [102] proposed a model that incorporates the restoration interdependencies among different infrastructure networks (e.g., water, power, transportation) into the network design and scheduling problems. They also investigated the effects of centralized and decentralized decision making on the service levels across infrastructures.

Xu et al. [128] proposed a stochastic integer program to find the optimal schedule for inspection, damage evaluation, and repair of post-earthquake damaged electric power system. Their aim was to minimize the average time that each customer is without power. Arab et al. [3] developed a mixed-integer model for preventive maintenance program in improving the reliability of electric power systems. Their model considers component deterioration, as well as two competing and independent failure modes including failure due to loss of reliability and failure due to hurricane damages. Their objective is to minimize the downtime cost of the power system due to component outage. They used a stochastic dynamic programming model to derive the optimal maintenance policy for the component.

Arif et al. [4] proposed a two-stage method for the outage management of power distribution systems. The first stage is to cluster repair tasks based on their distances from the depots and the availability of resources. The second stage is to co-optimize the repair, reconfiguration, and distributed generation dispatch considering routing repair crews to maximize the picked-up loads and minimize the repair time. By integrating infrastructure restoration with transportation network dispatch, Morshedlou et al. [79] proposed a new problem that addresses the dependent relationship between a disrupted infrastructure network and the routing network that connects all disrupted components.

In the context of maintenance vehicle routing, Van Hentenryck et al. [115] developed a joint model on how to schedule and route a fleet of repair crews to restore the power network as fast as possible after a disaster. Garcia et al. [33] proposed a mathematical model to schedule maintenance vehicles when considering emergency scenarios in electric distribution systems, from their corresponding GPS information to assign the most appropriate set of pending emergency orders previously defined. They must be assigned to the most appropriate set of pending emergency orders previously defined.

In the context of resource allocation for restoration of power systems, Yao and Min [129] presented three mathematical goal programming models in order to locate repair units and restore transmission and distribution lines in an efficient manner. The first model finds the optimal repair-unit dispatch tactical plan for a forecast of adverse weather conditions. The second model derives the optimal repair-unit location for a short-term strategic plan under normal weather conditions. The third model finds the optimal number of repair units for a long-term strategic plan. Coffrin et al. [15] developed a power system stochastic inventory model to stockpile components in order to recover from blackouts as best as possible after a disaster. Their proposed mixed-integer programming model combines power flow simulators, discrete storage decisions, discrete repair decisions given the storage decisions, and a collection of scenarios describing the potential effects of the disaster. In Wang et al., [117], a decision-making model was proposed to manage the required resources for economic power restoration operation. The optimal number of depots, the optimal location of depots, and the optimal number of repair crews were determined by their model in order to minimize the transportation cost associated with restoration operation. Arab et al. [2] proposed a stochastic resource allocation model for repair and restoration of potential damages to the power system infrastructure located on the path of an upcoming hurricane to minimize potential damages to power system components in a cost-effective manner.

Comprehensive surveys of models and algorithms for emergency response logistics in electric distribution systems, including reliability planning with fault considerations and contingency planning models, were presented in [87] and [88]. Borba et al. [7] also presented

the workforce problem focused on power utilities, as well as some topics related to power distribution planning at strategic, tactical and operational levels. Their survey reveals that most studies of the workforce problem in electrical distribution utilities analyses operational planning.

Although a variety of optimization models for power system planning for disruptive events have been addressed in the literature, to the best of my knowledge, in almost all above studies, the concentration is on the scheduling planning of repair crews and sequencing and routing for doing restoration. Knowledge about the state of the power network is used in their optimization models, and it is assumed that prior knowledge about the available number of repair crews or materials is available. However, in the case of extreme events, such as hurricanes, local crews are not enough for restoring power in reasonable time and utilities have to request crews from other states or sometimes other countries. These decisions should be made prior to the event because in most cases it takes time until external crews arrive and they significantly affect other decisions made by the utilities after the event. In this chapter, an efficient decision support tool is developed for proactive restoration planning of power systems to minimize the expected restoration costs, and maximize customer satisfaction by shortening the restoration period.

### 5.3 Problem Statement

This study focuses on developing an optimization approach based on MSSP for modeling a real-world crew coordination planning problem that many utility companies deal with in face of a natural hazard. The setting of this problem is derived based on my conversations with personnel from multiple utilities. It should be noted that the parameters of the model are not data driven, but based on some estimates obtained through these conversations. Thus, they should not be considered as parameters' actual values. However, by doing sensitivity analysis, I evaluate their influence on the decisions.

The proposed optimization approach represents a model to request crews from internal and external sources, establish staging areas, and allocate and relocate repair crews to different staging areas and districts (demand zones) in the face of extreme power outages caused by natural hazards. A planning horizon divided into multiple periods is taken into account to capture the variations of my network parameters and decisions. There are many side constraints and assumptions that are treated concurrently in this problem while minimizing the expected costs and maximizing the utility level of service (or customer satisfaction). In this section, I present the general description and assumptions of my problem. I shall call this problem the *repair crew coordination problem (RCCP)* . The essential assumptions of

this problem are as follows:

**Considering a network with three echelons:** A supply chain network is considered with three echelons: (i) operating companies and contracting agencies, (ii) staging areas and stores, and (iii) demand zones. In the first echelon, there exist several utility companies or contracting agencies that share and provide repair crews. Some of these utility companies work as parts of a large electric energy company. I call this set of companies the *local operating companies* (OCs) and denote them by  $\mathcal{J}$ , where each local OC is presented by  $j \in \mathcal{J}$ . Other companies and agencies that do not belong to this set are called *external companies*. If needed, local operating companies share their resources with each other with lower costs compared to the case in which resources are provided from external companies.

In the second echelon, utility stores and a set of potential locations (e.g., military facilities, college facilities) for establishing staging areas exist. Staging areas are closed and used for other purposes in normal periods. But, in case of an extreme event in which a large number of external crews are supposed to come for faster power restoration, these areas are opened. For each period, the model needs to decide which existing staging area to be closed and which potential one to be opened. The set of staging areas of OC  $j$  is denoted by  $\mathcal{L}(j)$ , where each area is presented by  $l \in \mathcal{L}(j)$ . Based on conversations with personnel from multiple utilities, it is assumed that each staging area has a specific capacity for holding and handling crews. It is presumed that all crews are staged in utility stores or staging areas before being sent to demand zones and thus, they cannot go directly from OCs to the demand zones.

Finally, each local OC serves an area, which is divided into districts with boundaries defined by the operating company serving the area. These districts are the problem's demand zones, which are the last echelon of my network. Because there exists one and only one store in each district, I use the same notation for both of them. Thus, the set of districts of OC  $j$  is also denoted by  $\mathcal{K}(j)$ , where each district is presented by  $k \in \mathcal{K}(j)$ . It is not mandatory to satisfy all crew needs in each period and some of it can be transferred to the next periods to be satisfied. However, this reduces customer satisfaction, which is one of my objectives to be maximized.

**Repair crew sharing between utilities in multiple stages:** Two types of crew sharing are considered in my model. Internal crew sharing is done inside the service territory of each OC and crews can move between stores and staging areas. External crew sharing is done between two local OCs or a local OC and an external company. Crews cannot directly be transshipped from an utility site in the service area of one local OC to a utility site within another local/external company. They should first be sent to their own local OC, then to the target local/external company and from there to one of its staging areas or stores. Figure 5.1 illustrates the crew supply network and decisions in my problem.

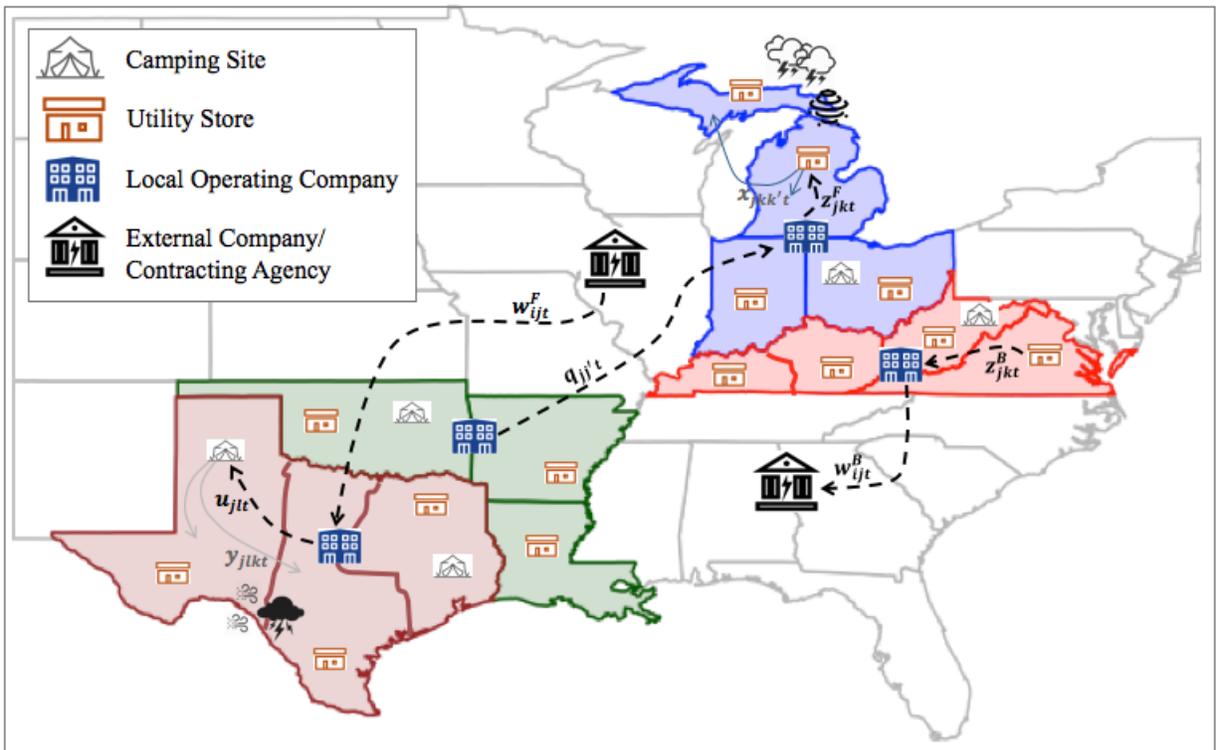


Figure 5.1: The crew supply network structure for power restoration in disasters. This figure demonstrates my network including utility companies, contracting agencies, staging areas, stores and demand points. It also shows resource flows between these entities for my repair crew coordination problem.

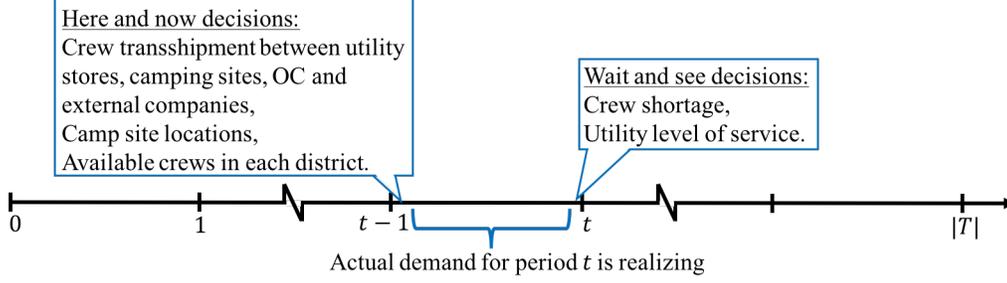


Figure 5.2: Decision-making process in my optimization model include two types of variables: (i) here and now decisions and (ii) wait and see decisions.

**Modeling different types of decisions:** In the proposed network, four types of decisions are required to be determined: location, transportation, shortage and utility levels. Location decisions include opening a staging area or closing an existing one. Transportation decisions include decisions about the number of crews transshipped between various facilities. Shortage is the difference between the needed demand and the satisfied demand, and utility levels represent customer satisfaction and are a function of total demand and satisfied demand. Each of these decisions is made in multiple stages and in every stage, some decisions are made based on the past realizations of parameters and some based on the forecasts for the future. Accordingly, decisions are divided into two categories, here-and-now (HN) decisions and wait-and-see (WS) decisions. The decisions regarding the HN variables have to be made at the beginning of each period, and we cannot wait until we have more or full information on the uncertain parameters in that period. However, the WS decisions can be determined at the end of the period, after the value of uncertain parameters in that period is revealed. In my problem, location and transportation decisions are HN type and have to be set before knowing stochastic demands at each period. Shortage and utility decisions on the other side are WS decisions and will be made only after the demands are disclosed in each period. Figure 5.2 demonstrates how these decisions should be determined over the planning horizon. Set  $\mathcal{T}$  represents the periods within a planning horizon,  $t \in \mathcal{T}$ , and  $t = 1, 2, \dots, |\mathcal{T}|$ .

**Considering release time and transportation time for crew sharing:** Each OC or contracting agency has a specific release time for its crews, and when they send their crews to other OCs or contracting agencies, their crews have to stay there until their release period ends. During this period, they can relocate between the stores and staging areas of that OC. Release times of a crew sent from OC  $j$  or company/contracting agency  $i$  are respectively represented by  $\rho_j^J$  and  $\rho_i^I$ . I also consider transportation time when crews are sent to somewhere outside their own operating company. Transit times between OC  $j$  and OC  $j'$  and the company/contracting agency  $i \in \mathcal{I}$  are respectively represented by  $\tau_{jj'}^J$ , by and

$\tau_{ij}^I$ . During these periods, crews are out of service.

**Coordinating "heavy" crew type:** Basically, there are two types of crews (i) big/heavy crew and (ii) service/light crew. Heavy crews are usually 4-7 people with a heavier equipment and they are able to repair any type of damage to the power system (e.g., setting poles up). Service/light crew on the other side are generally 2 people with a bucket truck and they can handle smaller tasks. In my problem, I propose a coordination model for the "heavy" crew type.

**Having two contradicting objectives:** I consider two objectives: (i) minimizing costs and (ii) maximizing customer satisfaction. Total costs include administration costs of crews and the cost of establishing staging areas, where administration costs of crews includes transportation cost and daily salary of crews. Customer satisfaction is represented using a utility function, which is a function of two variables: available crews in a demand area, and extra crew need for repairing the remaining damages. To deal with the nonlinearity arises from the utility function, I employ the triangle method [65] and convert the the non-linear utilities to a linear approximation form. The two objectives are optimized simultaneously. To convert my bi-objective problem into a single objective counterpart, I employ the improved version of the  $\epsilon$ -constraint method, namely AUGMECON [73].

**Uncertainty in crew needs:** Due to variability in the impact of hazards on the power system and restoration time, there exists inherent uncertainty in the number of repair crews needed for restoring power. Despite the existing uncertainty, I can estimate the future crew needs in the form of probability distributions. Using historical data (i.e., data presented in Chapter IV) as well as a probabilistic machine learning model (i.e., model presented in Chapter III [52]), I develop a statistical model predicting the damage to power system components including overhead and underground conductors, overhead and underground transformers, and poles. Before training the models, I re-frame my time series dataset with a window width of one. This means that I use the previous time step values of damages as new features. The probabilistic power system damage predictions are produced for every period in the planning horizon based on the forecasted weather and other influencing factors. These predicted probability distributions are then converted to the probability distributions of need for crews using data in Table 5.4, which is obtained from interviews with utility personnel.

The resulting continuous probability distributions for crew needs over the planning horizon form a multidimensional stochastic process. This stochastic process is approximated (discretized) via a set of discrete scenarios and multi-stage stochastic program (MSSP) is used to formulate my dynamic decision model. To generate scenarios, I employ the Latin Hypercube Sampling (LHS), introduced by Olsson et al. [84]. Compared with the Monte

Carlo simulation method, LHS can approximate the stochastic process through fewer sampling iterations while cover more of the domain of the random variables [59, 54]. Despite the use of an efficient algorithm, the scenario generation process for my multi-echelon and multi-period problem results in a very large number of scenarios, which makes the MSSP model hard to solve. Therefore, it is necessary to efficiently decrease the number of scenarios. For scenario reduction, I employ the backward reduction technique presented by Dupačová et al.[24]. Finally, the output scenarios is converted into a scenario tree using the forward scenario tree construction method [46].

**Multi-stage stochastic programming model:** A multi-stage stochastic program (MSSP) allows us to have several decision layers, where random outcomes are progressively realized, and the crew transshipment decisions should be adapted to this process. In general, a  $T + 1$ -stage stochastic program includes a sequence of random parameters  $\xi_1, \xi_2, \dots, \xi_T$  with a discrete support. A scenario is a realization of these stochastic parameters, and a scenario tree represents the progressive observation of random parameters. To model stochasticity in the number of crew needs as a scenario tree, a set of scenarios  $\mathcal{S}$  with a countable size  $S = |\mathcal{S}|$  is defined. The corresponding scenarios' probabilities are  $\pi_1, \pi_2, \dots, \pi_S$ , and a realization of the stochastic parameters for scenario  $s \in \mathcal{S}$  is presented by  $(\xi_1^s, \xi_2^s, \dots, \xi_T^s)$ , where  $\xi_t^s = \{d_{jkt} : j \in \mathcal{J}, k \in \mathcal{K}(j)\}$  is a realization for the number of crew needs on period  $t \in \mathcal{T}$  over different districts under scenario  $s \in \mathcal{S}$ .

Note that the realization of random parameters  $\xi_1, \xi_2, \dots, \xi_{t-1}$  has been observed at intermediate stage  $t$ , and the remaining uncertainty includes  $\xi_t, \xi_{t+1}, \dots, \xi_T$ . In a MSSP, a policy should be *non-anticipative*, meaning that the decisions made at stage  $t$  must not be dependent on the future realization of stochastic parameters. There are two common ways for formulating a MSSP ([23]). In the first, a MSSP is formulated as a sequence of nested two-stage stochastic programs in which non-anticipativity is implicitly imposed. In the second (used in this chapter), a set of *non-anticipativity constraints* (NAC) is explicitly modeled.

Figure 5.3 (left-hand side) shows an example of a scenario tree with four stages and eight scenarios. Figure 5.3 (right-hand side) is an alternative representation of the scenario tree, which is called *scenario fan*, where the individual scenarios observed in the particular stages are aggregated over all periods to form eight scenarios. However, this scenario fan is *not permissible*. If I solve my problem for each of the scenarios, the solution found might not be feasible for the overall problem because they imply decisions that anticipate future uncertain events. So, I need to enforce NACs to have permissible decisions. The dashed ovals covering the nodes present NACs. These constraints assure that all the decisions in a given stage  $t$  are identical for each pair  $(s, s')$  of scenarios with a common ancestor node in that stage. If two scenarios  $s$  and  $s'$  share the same history of random parameters  $\xi^s$  and  $\xi^{s'}$  up to stage

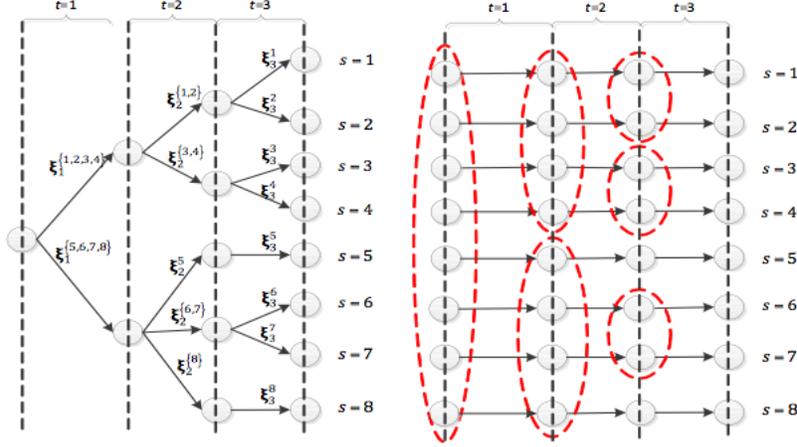


Figure 5.3: A scenario tree example for a MSSP with four stages and eight scenarios (left figure) and the decomposed scenario tree to represent the non-anticipativity constraints (right figure).

$t$ , then the decisions made at stage  $t$  are the same among all scenarios placed in the same scenario bundle. For example, since all eight scenarios have the same realizations at stage 1, they share the same scenario bundle, and so a NAC is imposed to guarantee that the same crew transshipment decisions are made at all nodes in this scenario bundle.

## 5.4 Mathematical formulation

In this section, the MSSP model for the repair crew coordination problem described in Section 5.3 is presented. In Section 5.4.1, I develop a MSSP model which has some non-linear terms. In Section 5.4.2, I linearize these terms in the original model and convert the bi-objective problem into a single objective counterpart using  $\epsilon$ -greedy methodology.

### 5.4.1 Multi-stage Stochastic Mixed-Integer Program Model

The planning horizon  $\mathcal{T}$  is defined for my problem.  $\mathcal{T}$  is the set of periods from current period until period  $T = |\mathcal{T}|$ , over which we decide the number of crews that should be transshipped between different facilities, and whether or not to establish staging areas. The other indices, parameters and decisions used in the model are given in Table 5.1.

**Objective function for costs:** The objective function (5.1) minimizes the sum of the administration costs of crews (i.e., terms 1, 2, 3, 4, 5 and 6 respectively) and the costs of establishing and keeping open staging areas (i.e., last two terms). The first term includes transportation cost and daily salary of crews sent from store  $k$  to district  $k'$ . The second term includes transportation cost of crews transshipped from staging area  $l$  to district  $k$ . The

Table 5.1: The description of indices, parameters and decisions of the model.

<b>Sets and Indices</b>	
$\mathcal{T}$	The set of periods indexed by $t, t' \in \mathcal{T}$ .
$\mathcal{J}$	The set of operating companies (OC) indices, $j, j' \in \mathcal{J}$ .
$\mathcal{K}(j)$	The set of districts/stores of OC $j \in \mathcal{J}$ indexed by $k \in \mathcal{K}(j)$ .
$\mathcal{L}(j)$	The set of staging areas of OC $j \in \mathcal{J}$ indexed by $l \in \mathcal{L}(j)$ .
$\mathcal{I}$	The set of other companies and contracting agencies indexed by, $i \in \mathcal{I}$ .
$\mathcal{S}$	The set of scenario indexed by, $s \in \mathcal{S}$ .
<b>Parameters</b>	
$V_{jlt}$	Total external crew capacity of staging area $l \in \mathcal{L}(j)$ at period $t \in \mathcal{T}$ .
$U_{jkt}$	Total external crew capacity of store $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ .
$N_{jkt}^K$	Total number of available internal crews in store $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ .
$N_{it}^I$	Total number of available crews in the company/contracting agency $i \in \mathcal{I}$ at period $t \in \mathcal{T}$ that we can request for.
$a_{jkt}$	Proportion of crews in store $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ that can be transshipped to other OCs.
$d_{jkts}$	The crew demand at district $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$\pi_s$	Probability of occurrence of scenario $s \in \mathcal{S}$ .
$Ed_{it}$	The crew request by the company/contracting agency $i \in \mathcal{I}$ to OC $j \in \mathcal{J}$ at period $t \in \mathcal{T}$ .
$\rho_j^J$	Release time of a crew sent from OC $j$ to outside the OC.
$\rho_i^I$	Release time of a crew sent from the company/contracting agency $i \in \mathcal{I}$ to OCs.
$\tau_{jj'}$	Transit time between OC $j$ and $j'$ ( $j, j' \in \mathcal{J}$ ).
$\tau_{ij}^I$	Transit time between OC $j \in \mathcal{J}$ and the company/contracting agency $i \in \mathcal{I}$ .
$FC_{jlt}$	Fixed cost of establishing staging area $l \in \mathcal{L}(j)$ at period $t \in \mathcal{T}$ .
$AC_{jlt}^{JL}$	Administration cost of keeping staging area $l \in \mathcal{L}(j)$ at period $t \in \mathcal{T}$ open.
$AC_{jkk't}^K$	Administration cost of a crew sent from site $k$ to district $k'$ ( $k, k' \in \mathcal{K}(j)$ ) at period $t \in \mathcal{T}$ .
$AC_{jkt}^L$	Administration cost of a crew sent from staging area $l \in \mathcal{L}(j)$ to district $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ .
$AC_{jkt}^{JK}$	Administration cost of a crew transshipped between site $k \in \mathcal{K}(j)$ and OC $j \in \mathcal{J}$ at period $t \in \mathcal{T}$ .
$AC_{jlt}^{JL}$	Administration cost of a crew transshipped between staging area $l \in \mathcal{L}(j)$ and OC $j \in \mathcal{J}$ at period $t \in \mathcal{T}$ .
$AC_{jj't}^{JJ}$	Administration cost of a crew sent from OC $j$ to $j'$ ( $j, j' \in \mathcal{J}$ ) at period $t \in \mathcal{T}$ .
$AC_{ijt}^{JI}$	Administration cost of a crew sent from the company/contracting agency $i \in \mathcal{I}$ to OC $j \in \mathcal{J}$ at period $t \in \mathcal{T}$ .
<b>Here and Now Decisions</b>	
$x_{jkk'ts}$	The number of crews sent from store $k$ to district $k'$ ( $k, k' \in \mathcal{K}(j)$ ) at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$y_{jlkts}$	The number of crews sent from staging area $l \in \mathcal{L}(j)$ to district $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$u_{jlt}$	The number of crews sent from OC $j \in \mathcal{J}$ to staging area $l \in \mathcal{L}(j)$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$z_{jkt}^F$	The number of crews sent from OC $j \in \mathcal{J}$ to store $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$z_{jkt}^B$	The number of crews sent from store $k \in \mathcal{K}(j)$ to OC $j \in \mathcal{J}$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$q_{jj't}$	The number of crews sent from OC $j$ to $j'$ ( $j, j' \in \mathcal{J}$ ) at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$w_{ijts}^F$	The number of crews sent from the company/contracting agency $i \in \mathcal{I}$ to OC $j \in \mathcal{J}$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$w_{ijts}^B$	The number of crews sent from OC $j \in \mathcal{J}$ to the company/contracting agency $i \in \mathcal{I}$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$v_{jlt}$	Binary variable equal to 1 if staging area $l \in \mathcal{L}(j)$ within the service territory of OC $j \in \mathcal{J}$ is open at $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$v_{jlt}^E$	Binary variable equal to 1 if staging area $l \in \mathcal{L}(j)$ within the service territory of OC $j \in \mathcal{J}$ is established at $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$o_{jkts}$	The number of available crews (internal and external) in district $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
<b>Wait and See (WS) Decisions</b>	
$\Delta_{jkts}$	The number of crew shortage in district $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .
$\varphi_{jkts}$	Utility level of service in district $k \in \mathcal{K}(j)$ at period $t \in \mathcal{T}$ under scenario $s \in \mathcal{S}$ .

third term represents total costs of transportation and salary of crews for  $\rho_j^J$  days. The fourth and fifth terms represent the transportation cost of crews transshipped between OC  $j$  and utility store  $k$  or staging area  $l$ . Finally, the sixth term represents total costs of transportation between OC  $j$  and the company/contracting agency  $i$  and the salary of crews for  $\rho_i^I$  days.

$$(5.1) \quad \min \quad OF1 = \sum_{j,k,k',t} x_{jkk'ts} AC_{jkk't}^K + \sum_{j,l,k,t} y_{jlkts} AC_{jlkts}^L + \sum_{j,j',t} q_{jj'ts} AC_{jj't}^{JJ} + \sum_{j,k,t} u_{jlt} AC_{jlt}^{JL} + \sum_{j,k,t} (z_{jkt}^F + z_{jkt}^B) AC_{jkt}^{JK} + \sum_{i,j,t} w_{ijts} AC_{ijts}^{JI} + \sum_{j,l,t} v_{jlt}^E FC_{jlt} + \sum_{j,l,t} v_{jlt} AC_{jlt}$$

**Objective function for utility level of service:** The objective function (5.2) is to maximize the utility level of service in all districts. This utility level measure is calculated through constraint (5.16) for each district in each period and scenario.

$$(5.2) \quad \max \quad OF2 = \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}(j)} \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{S}} \pi_s \varphi_{jkt}$$

**Constraints for the relationship between flows and location capacities:** Constraints (5.3) state that if no staging area is established at a location, there should not be any flow from the OC to this staging area; otherwise, flow from the OC to this site should be less than its capacity. Constraints (5.4) ensure that any flow from a staging area to other districts should be less than the total crews that are staged in the staging area. Constraints (5.5) state that the number of external crews sent from the OC to the stores should be less than their capacity. Constraints (5.6) impose some capacity restriction on the number of crews sent from the stores to their OCs to be transshipped to other OCs. Constraints (5.7) ensure that the number of crews sent from a store to other districts or OCs should be less than available crews in this site. Finally, constraints (5.8) impose capacity restriction on the number of crews received from the external companies or contracting agencies.

$$(5.3) \quad u_{jlt} \leq V_{jlt} * v_{jlt} \quad , \quad \forall j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.4) \quad \sum_{k \in \mathcal{K}(j)} y_{jlkts} \leq u_{jlt} \quad , \quad \forall j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.5) \quad z_{jkt}^F \leq U_{jkt} \quad , \quad \forall j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.6) \quad \sum_{t'=t-\rho_j}^t z_{jkt's}^B \leq N_{jkt}^K * a_{jkt} \quad , \quad \forall j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.7) \quad \sum_{t'=t-\rho_j^J}^t z_{jkt's}^B + \sum_{k' \in \mathcal{K}(j)} x_{jkk'ts} \leq z_{jkt's}^F + N_{jkt}^K, \quad \forall j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.8) \quad \sum_{t'=t-\rho_i}^t w_{ijt's}^F \leq N_{it}^I, \quad \forall i \in \mathcal{I}, j \in \mathcal{J}, t \in \mathcal{T}, s \in \mathcal{S}.$$

**Flow constraints:** Constraints (5.9) state that the number of available crews in each district is equal to the sum of the number of crews sent from stores and staging areas to that district. This shows that crews cannot be sent from OCs. They should first be staged in the stores or staging areas and from there transshipped to districts to do restoration. Constraints (5.10) ensure that every day, all the crews that are sent from stores to their OC are dispatched to other OCs or external company/contracting agencies. Constraints (5.11) state that at each period, the number of crews that are sent from other OCs, and other utility companies/contracting agencies to an OC is equal to the number crews that are sent from that OC to the staging areas and stores in their territory. When crews are dispatched to an OC, they stay there for a period with the length of their release time. But, they can relocate between the stores and staging areas of that OC. Constraints (5.11) show that each OC should plan for staging all the crews that are sent to its company for all their release periods.

$$(5.9) \quad o_{jkt's} \leq \sum_{k' \in \mathcal{K}(j)} x_{jk'k'ts} + \sum_{l \in \mathcal{L}(j)} y_{jlkt's}, \quad \forall j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.10) \quad \sum_{k \in \mathcal{K}(j)} z_{jkt's}^B = \sum_{j' \in \mathcal{J}} q_{jj'ts} + \sum_{i \in \mathcal{I}} w_{ijts}^B, \quad \forall j \in \mathcal{J}, t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.11) \quad \sum_{j' \in \mathcal{J}} \sum_{t'=t-\rho_{j'}^J, -\tau_{jj'}^J}^{t-\tau_{jj'}^J} q_{j'jt's} + \sum_{i \in \mathcal{I}} \sum_{t'=t-\rho_i^I, -\tau_{ij}^I}^{t-\tau_{ij}^I} w_{ijts}^F = \sum_{l \in \mathcal{L}(j)} u_{jlts} + \sum_{k \in \mathcal{K}(j)} z_{jkt's}^F, \quad \forall j \in \mathcal{J}, t \in \mathcal{T}, s \in \mathcal{S}.$$

**Constraints related to establishing staging areas:** Constraints (5.12) ensure that if a staging area is closed,  $v_{jlt's}^E$  variable is zero. Constraints (5.13), on the others side, ensure that  $v_{jlt's}^E$  variable is set equal to one at the period it is established.

$$(5.12) \quad v_{jlt's}^E \leq v_{jlts}, \quad \forall j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.13) \quad v_{jlt's} - v_{jl(t-1)s} \leq v_{jlt's}^E, \quad \forall j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

**Demand and shortage constraints:** External utility companies/contracting agencies may also have crew requests from the local OCs. Usually, companies are committed to satisfy their demands as many as possible. But, costs associated with these movements should be paid by the utility company/contracting agency that requested for the crews. Thus, I include these requests in my model by adding some constraints. Accordingly, constraints (5.14) are added to assure that these demands are met. Constraints (5.15) assert that the sum of satisfied demand and not-satisfied demand of each district in each period should be equal to total crew need in this demand point, which is equal to the demand at the current period plus not-satisfied demand from all previous periods.

$$(5.14) \quad \sum_{j \in \mathcal{J}} \sum_{t'=t-\rho_j^I-\tau_{ij}^I}^{t-\tau_{ij}^I} w_{ijts}^B \geq Ed_{it} \quad , \quad \forall i \in \mathcal{I}, t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.15) \quad o_{jkts} + \Delta_{jkts} \geq d_{jkts} + \Delta_{jk(t-1)s} \quad , \quad \forall j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

**Utility constraint:** Constraints (5.16) compute the utility level of serving a district in each time period. In next section, I will explain how these utility functions are defined and how I can convert these nonlinear constraints (5.16) into linear ones.

$$(5.16) \quad \varphi_{jkts} = f\left(\frac{O_{jkts}}{O_{jkts} + \Delta_{jkts}}\right) \quad , \quad \forall j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

**Non-anticipativity constraints:** Constraints (5.17-5.27) represent the NAC for my MSSP model where  $\xi_{[t]}^s = (\xi_1^s, \dots, \xi_{t-1}^s)$ . All the decisions represented in these constraints have to be determined before knowing the potential crew need at each stage. Indeed, the NAC are presumed only for HN decisions ([56]).

$$(5.17) \quad x_{jkk'ts} = x_{jkk'ts'} \quad , \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, k, k' \in \mathcal{K}(j), t \in \mathcal{T}.$$

$$(5.18) \quad y_{jlkts} = y_{jlkts'} \quad , \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, k \in \mathcal{K}(j), l \in \mathcal{L}(j), t \in \mathcal{T}.$$

$$(5.19) \quad u_{jllts} = u_{jllts'} \quad , \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}.$$

$$(5.20) \quad z_{jkts}^F = z_{jkts'}^F \quad , \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}.$$

$$(5.21) \quad z_{jkts}^B = z_{jkts'}^B \quad , \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}.$$

$$(5.22) \quad q_{jj'ts} = q_{jj'ts'} \quad , \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j, j' \in \mathcal{J}, t \in \mathcal{T}.$$

$$(5.23) \quad w_{ijts}^F = w_{ijts'}^F \quad , \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, i \in \mathcal{I}, t \in \mathcal{T}.$$

$$(5.24) \quad w_{ijts}^B = w_{ijts'}^B \quad , \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, i \in \mathcal{I}, t \in \mathcal{T}.$$

$$(5.25) \quad v_{jllts} = v_{jllts'} \quad , \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}.$$

$$(5.26) \quad v_{j l t s}^E = v_{j l t s'}^E, \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}.$$

$$(5.27) \quad o_{j k t s} = o_{j k t s'}, \quad \forall s, s' \in \mathcal{S} : \xi_{[t]}^s = \xi_{[t]}^{s'}, j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}.$$

### 5.4.2 Linear MSSP Crew Coordination Model

**Linear approximation of the utility function:** The term  $f(\frac{o_{j k t s}}{o_{j k t s} + \Delta_{j k t s}})$  introduces a non-linearity into the objective functions (5.2). Generally, non-linear programming models are more difficult to solve than linear ones. Hence, I convert the presented non-linear model to a linear approximation form through a piecewise linear approximation technique. Since  $f(\frac{o_{j k t s}}{o_{j k t s} + \Delta_{j k t s}})$  is a function of two variables, I employ the triangle method [65], which is an extension of the piecewise linear approximation method of single variable functions to the two-variable case.

In this case, the piecewise linear approximation is obtained by introducing  $n$  sampling points  $o_{j k t s}^1, \dots, o_{j k t s}^n$  on the  $o_{j k t s}$  variable, and  $m$  sampling points  $\Delta_{j k t s}^1, \dots, \Delta_{j k t s}^m$  on the  $\Delta_{j k t s}$  variable, with  $o_{j k t s}^1$  and  $o_{j k t s}^n$  (resp.  $\Delta_{j k t s}^1$  and  $\Delta_{j k t s}^m$ ) coinciding with the left and right extremes of the  $o_{j k t s}$  (resp.  $\Delta_{j k t s}$ ). The function  $f(\frac{o_{j k t s}}{o_{j k t s} + \Delta_{j k t s}})$  is evaluated for each breakpoint  $(o_{j k t s}^g, \Delta_{j k t s}^h)$  ( $g = 1, \dots, n; h = 1, \dots, m$ ). For any given  $(o_{j k t s}, \Delta_{j k t s})$  point, say  $(\bar{o}_{j k t s}, \bar{\Delta}_{j k t s})$ , with  $o_{j k t s}^g \leq \bar{o}_{j k t s} \leq o_{j k t s}^{g+1}$  and  $\Delta_{j k t s}^h \leq \bar{\Delta}_{j k t s} \leq \Delta_{j k t s}^{h+1}$ , let us consider the rectangle of vertices  $(o_{j k t s}^g, \Delta_{j k t s}^h), (o_{j k t s}^{g+1}, \Delta_{j k t s}^h), (o_{j k t s}^g, \Delta_{j k t s}^{h+1}), (o_{j k t s}^{g+1}, \Delta_{j k t s}^{h+1})$  and the two triangles produced by its diagonal  $[(o_{j k t s}^g, \Delta_{j k t s}^h), (o_{j k t s}^{g+1}, \Delta_{j k t s}^{h+1})]$ . (The triangles produced by the other diagonal could equivalently be used.) The function value is then approximated by a convex combination of the function values evaluated at the vertices of the triangle containing  $(\bar{o}_{j k t s}, \bar{\Delta}_{j k t s})$ .

In order to implement the above technique in my problem, it is necessary to include in the model, the variables and constraints that force any  $(o_{j k t s}, \Delta_{j k t s})$  point to be associated with the proper triangle surrounding it. Let us introduce  $nm$  continuous variables  $\alpha_{j k t s}^{gh}$  for each breakpoint  $g$  and  $h$ , such that  $\alpha_{j k t s}^{gh} \in [0, 1]$  ( $g \in \{1, \dots, n\}, h \in \{1, \dots, m\}$ ). Let  $hu_{j k t s}^{gh}$  and  $hl_{j k t s}^{gh}$  be two binary variables, respectively, associated with the upper and lower triangle in the rectangle of vertices  $(o_{j k t s}^g, \Delta_{j k t s}^h), (o_{j k t s}^{g+1}, \Delta_{j k t s}^h), (o_{j k t s}^g, \Delta_{j k t s}^{h+1}), (o_{j k t s}^{g+1}, \Delta_{j k t s}^{h+1})$ , with dummy values  $hu_{j k t s}^{0*} = hu_{j k t s}^{*0} = hu_{j k t s}^{n*} = hu_{j k t s}^{*m} = hl_{j k t s}^{0*} = hl_{j k t s}^{*0} = hl_{j k t s}^{n*} = hl_{j k t s}^{*m} = 0$ . Then, the convex combinations in the three-dimensional space is computed as follows:

$$(5.28) \quad \sum_{g=1}^n \sum_{h=1}^m \alpha_{j k t s}^{gh} = 1, \quad \forall j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.29) \quad o_{j k t s} = \sum_{g=1}^n \sum_{h=1}^m \alpha_{j k t s}^{gh} o_{j k t s}^g, \quad \forall j \in \mathcal{J}, k, k' \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.30) \quad \Delta_{j k t s} = \sum_{g=1}^n \sum_{h=1}^m \alpha_{j k t s}^{gh} \Delta_{j k t s}^h, \quad \forall j \in \mathcal{J}, k, k' \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

Constraints (5.31) impose that, among all triangles, only one is used for the convex combination. Then, constraints (5.32) impose that the only  $\alpha_{j k t s}^{gh}$  values different from 0 can be those associated with the three vertices of such triangle. Constraints (5.33) calculate then the correct computation of the approximate value for  $\varphi_{j k t s}$ .

$$(5.31) \quad \sum_{g=1}^{n-1} \sum_{h=1}^{m-1} (hu_{j k t s}^{gh} + hl_{j k t s}^{gh}) = 1, \quad \forall j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.32) \quad \alpha_{j k t s}^{gh} \leq hu_{j k t s}^{gh} + hl_{j k t s}^{gh} + hu_{j k t s}^{g,h-1} + hl_{j k t s}^{g-1,h-1} + hu_{j k t s}^{g-1,h-1} + hl_{j k t s}^{g-1,h}, \quad \forall g \in \{1, \dots, n\}, h \in \{1, \dots, m\}, j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

$$(5.33) \quad \varphi_{j k t s} = \sum_{g=1}^n \sum_{h=1}^m \alpha_{j k t s}^{gh} f \left( \frac{O_{j k t s}^g}{O_{j k t s}^g + \Delta_{j k t s}^h} \right), \quad \forall j \in \mathcal{J}, k, k' \in \mathcal{K}(j), t \in \mathcal{T}, s \in \mathcal{S}.$$

**Single objective counterpart of the multi-objective model:** After obtaining a linear approximation for the model by applying the piecewise linear approximation technique, the problem is converted to a multi-objective mixed integer linear mathematical model. Several methods have been developed in the literature to tackle the multi-objective mathematical models such as the weighted sum,  $\epsilon$ -constraint, Chebycheff-based methods or the fuzzy programming. In this work, I employ the improved version of the  $\epsilon$ -constraint method, namely AUGMECON [73], by which the multi-objective model is converted to a single objective counterpart. This method is an appropriate approach for my problem, because it can handle non-convex pareto-optimal set, and assures the exact Pareto set with an efficient amount of computational effort.

It is well known that the  $\epsilon$ -constraint method has certain advantages in relation to the weighting method [73]. AUGMECON addresses some drawbacks of the conventional  $\epsilon$ -constraint method, namely, the guarantee of Pareto optimality of the obtained solution in the payoff table as well as in the generation process and the increased solution time for problems with several objective functions.

The formulation of the AUGMECON method for my problem is as follows:

$$(5.34) \quad \max \quad OF2 + eps \times \left(\frac{\beta}{b}\right)$$

*s.t.*

Constraints (5.3) – (5.15) and (5.17) – (5.33)

$$(5.35) \quad OF1 - \beta = \epsilon$$

$$(5.36) \quad \beta < 0$$

where  $\epsilon$  is the right hand-side parameter for the specific iteration drawn from the grid points of the objective functions 1. The Pareto-optimal solutions of the model are achieved by parametrical variation of this parameter. The parameter  $b$  is the range of the respective objective functions.  $\beta$  is the surplus variables of the respective constraint and  $eps \in [10^{-3}, 10^{-6}]$ . The range of  $\epsilon$  can be calculated by optimizing the constrained objective functions  $OF1$  separately subject to the constraints and constructing the pay-off table. Afterward, different values for  $\epsilon$  can be calculated by dividing the range of constrained objectives  $OF1$  to  $p$  equal intervals. Thus, I have in total  $p + 1$  grid points that are used to vary parametrically the right hand side of the  $OF1$ .

$$b = OF1^{max} - OF1^{min}; \quad \epsilon^m = OF1^{max} - \frac{b}{p} \times m; \quad m = 0, \dots, p - 1$$

In the multi-objective integer programming, the  $\epsilon$ -constraint method can be used to produce the exact (or complete) Pareto optimal set. In this case, the size of the Pareto set is finite and the AUGMECON is therefore suitable for generating the exact Pareto set.

## 5.5 Computational results and analysis

This section is presented in two main parts. First, the efficiency of the proposed MSSP for the repair crew coordination problem is investigated by comparing the value of solution obtained by MSSP and the one achieved by the two-stage version of the stochastic problem. In the second part, the application of the proposed model is discussed by solving one instance of the repair crew coordination problem based on the case study. I also carry out sensitivity analysis on the crucial parameters of my model.

### 5.5.1 Assessing the performance of the MSSP

To evaluate the presented model's performance, several test problems from different classes including small-sized, medium-sized and large-sized are created. The complete char-

acteristics of these test problems are provided in Table 5.2. The model is coded in Python 3.7.3 and solved using Gurobi 9.0.2. The computations are performed on a personal computer with Core i7-1600 MHz CPU, 3 GHz and 16 GB of RAM.

It should be noted that given the concentration of this study on problem formulation, solution interpretation and practical side of the study, rounding approach (for non-binary integer variables) is utilized to efficiently solve the mixed-integer programming problem. In this approach, the problem is solved as a linear program (LP) with continuous variables. Then, solutions are rounded to an integer one by searching out satisfactory solutions wherein the variable values are adjusted to nearby larger or smaller integer values. Finally, rounded solutions are compared with unrounded ones to ensure that all the constraints are still satisfied, and the difference between the optimal solutions obtained from solving the LP problem and the rounded solutions is not significant.

**The relative value of the multi-stage stochastic program:** To highlight the importance of my multi-stage stochastic model for the repair crew coordination problem, I compare the multi-stage stochastic model with the two-stage stochastic version of this problem. To this aim, I need to calculate the relative value of MSSP with respect to TSSP, which is obtained as follows [48]:

$$RVMS = \frac{obj^{MSSP} - obj^{TSSP}}{obj^{TSSP}} \times 100\%,$$

where  $obj^{MSSP}$  and  $obj^{TSSP}$  are respectively the optimal objective function values of multi-stage and two-stage stochastic formulations of my problem.

In order to convert my proposed MSSP model into a two-stage stochastic program (TSSP) model, I need to substitute constraints (5.37)-(5.47) with the non-anticipativity constraints (5.17)-(5.27) in the proposed MSSP model.

$$(5.37) \quad x_{jkk'ts} = x_{jkk'ts'} \quad , \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, k, k' \in \mathcal{K}(j), t \in \mathcal{T}.$$

$$(5.38) \quad y_{jlkts} = y_{jlkts'} \quad , \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, k \in \mathcal{K}(j), l \in \mathcal{L}(j), t \in \mathcal{T}.$$

$$(5.39) \quad u_{jltts} = u_{jltts'} \quad , \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}.$$

$$(5.40) \quad z_{jks}^F = z_{jks'}^F \quad , \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}.$$

$$(5.41) \quad z_{jks}^B = z_{jks'}^B \quad , \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}.$$

$$(5.42) \quad q_{jj'ts} = q_{jj'ts'} \quad , \quad \forall s, s' \in \mathcal{S}, j, j' \in \mathcal{J}, t \in \mathcal{T}.$$

$$(5.43) \quad w_{ijts}^F = w_{ijts'}^F \quad , \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, i \in \mathcal{I}, t \in \mathcal{T}.$$

$$(5.44) \quad w_{ijts}^B = w_{ijts'}^B \quad , \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, i \in \mathcal{I}, t \in \mathcal{T}.$$

$$(5.45) \quad v_{jltts} = v_{jltts'} \quad , \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}.$$

Table 5.2: Results from solving test problems with Gurobi solver and estimated values of the multi-stage stochastic program.

Test problems	$ \mathcal{S} $	$ \mathcal{T} $	$ \mathcal{J} $	$ \mathcal{I} $	$  \mathcal{K}(j)  $	$  \mathcal{L}(j)  $	Average RVMS
Test problem 1	9	5	2	1	[3,4]	[2,4]	3.04%
Test problem 2	9	7	3	2	[4,4,3]	[3,4,2]	2.14%
Test problem 3	17	9	5	3	[5,2,3,6,3]	[3,2,3,4,2]	6.32%
Test problem 4	29	10	7	4	[5,4,3,6,3,6,5]	[3,2,1,3,1,3,4]	4.30%
Test problem 5	41	14	10	5	[5,4,3,6,3,6,5,3,6,5]	[4,2,1,3,1,3,3,3,6,5]	3.49%

$$(5.46) \quad v_{jlt s}^E = v_{jlt s'}^E, \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, l \in \mathcal{L}(j), t \in \mathcal{T}.$$

$$(5.47) \quad o_{jkt s} = o_{jkt s'}, \quad \forall s, s' \in \mathcal{S}, j \in \mathcal{J}, k \in \mathcal{K}(j), t \in \mathcal{T}.$$

In my problem formulation, I have two objective functions and the value of the combined objective (5.34) depends on the chosen value for the parameter  $\epsilon$  in constraint (5.35). Thus, for 10 different values of  $\epsilon$ , I solve both the MSSP and TSSP problems and obtain  $obj^{TSSP}$ ,  $obj^{MSSP}$ , and accordingly calculate *RVMS*. The last column in Table 5.2 presents the average RVMS values for different test problems. As shown by Table 5.2, the relative superiority of the MSSP is verified in comparison with the TSSP. It should be highlighted that adding constraints (5.37)-(5.47) instead of NACs reduces the feasible region of the problem and therefore, the optimal solution of the TSSP becomes less than or equal to the optimal solution of the MSSP.

### 5.5.2 Application of the proposed stochastic model

In this section, an illustrative example of my model based on my case study is discussed. I develop this illustrative example based on the data from one of the utility companies I have been working with. In this example, crew transshipment decisions are made for seven local operating companies (i.e.,  $|\mathcal{J}| = 7$ ), where each one respectively serves  $||\mathcal{K}(j)|| = [5, 4, 3, 6, 3, 6, 5]$  districts. I develop my model for a  $|\mathcal{T}| = 10$  day planning horizon, and I assume that  $|\mathcal{I}| = 4$ ,  $||\mathcal{L}(j)|| = [3, 2, 1, 3, 1, 3, 4]$ , and the utility function is  $f(x) = \sqrt{x}$ . Other corresponding parameters are generated based on Table 5.3. Based on the procedure described in Section 5.3, I come up with a scenario tree with  $|\mathcal{S}| = 29$  scenarios. Using this example, I study the impact of four important parameters in the model. These parameters include (i) proportion of crews in each store that can be transshipped to other OCs ( $a_{jkt}$ ), (ii) the number of crews requested by the other company/contracting agency to get from local OCs ( $Ed_{it}$ ), (iii) the crew need at each district ( $d_{jkt s}$ ), and (iv) the utility function ( $f(\cdot)$  function in equation (5.16)).

Table 5.3: Characteristics of parameters for the test problems.

Symbol	Value	Symbol	Value	Symbol	Value
$V_{jlt}$	$U(100, 200)$	$\rho_j^J$	$U(2, 4)$	$AC_{jkk't}^K$	$U(15, 30)$
$U_{jkt}$	$U(5, 20)$	$\rho_i^I$	$U(2, 4)$	$AC_{jkt}^{JK}$	$U(150, 250)$
$N_{jkt}^K$	$U(20, 100)$	$\tau_{jj'}^J$	$U(0, 2)$	$AC_{jj't}^{JJ}$	$U(100, 2000)$
$N_{it}^I$	$U(100, 200)$	$\tau_{ij}^I$	$U(0, 2)$	$AC_{jlt}^{JL}$	$U(150, 200)$
$Ed_{it}$	$U(0, 40)$	$FC_{jlt}$	$U(1e + 06, 3e + 06)$	$AC_{jlt}^L$	$U(15, 30)$
		$AC_{jlt}$	$U(5e + 04, 2e + 05)$	$AC_{ijt}^{JI}$	$U(700, 2500)$

Since I used the augmented  $\epsilon$ -constraint method for combining the objectives, I have pareto optimal solutions instead of single objective values. By changing the  $\epsilon$  value in the constraint (5.35), different pairs of pareto optimal solutions are obtained that are shown with a curve in Figures 5.4 - 5.7. Figure 5.4 represents my analysis for three different values of parameter  $a$ . I solve my case study for 3 different values of parameter  $a$  while I keep other parameters of the model constant. This figure shows that as I increase parameter  $a$ , optimal objective values improve. In another analysis, I study the effect of change in two parameters including  $Ed$  and  $a$ , simultaneously. I solve my case study for 4 different scenarios. In each case, I double the value of one of these two parameters. Results are shown in Figure 5.5. We see that with increase in external demand ( $Ed$ ), optimal objective value significantly deteriorates. That is, same level of utility is obtained with much higher costs. We also see that in both levels of external demand, increase in parameter  $a$  improves the objective values, and this improvement is larger for the higher  $Ed$  values.

Based on my conversations with personnel of one of the utility companies with which I collaborated in this study, I found that they tend to take a very conservative approach in regard with the percentage of the crews that are allowed to be sent to other companies. In other words, they set a very low limit (i.e., 0.15 – 0.2) for the parameter  $a$ . This is because they believe that increasing this parameter may put their company at a higher risk of not being able to satisfy demand for crew needs in case of a sudden hazard. However, my results show that increasing parameter  $a$  not only does not deteriorate solutions, but also results in higher utility of service. What increases the risk for utility companies is the increase in external demand. Thus, a better alternative policy is to set no limit for parameter  $a$  (i.e.,  $a = 1$ ), but put limits on the maximum number of crews that all OCs together can provide

Table 5.4: Approximate crew hour need for fixing power system damages.

	Pole	Underground Transformers	Overhead Transformers	Underground Lines	Overhead Lines
Crew Need	4-8 hours	1 day	6-8 hours	1-2 days	1-2 hours

for external companies.

Moreover, I analyze sensitivity of my model with respect to increase in crew needs. Results are shown in Figure 5.6. Here, I have three scenarios for demand, including low, medium and high demand, where the medium demand is twice as big as the low demand and the high demand is four times as big as the low demand. As expected, we observe that given the fixed resources, as the demand increases, utility level of service decreases and also the system cost increases. Finally, I study the impact of different utility functions on model's optimal solutions. The three functions I considered are  $f(x) = \sqrt{x}$ ,  $f(x) = x$ , and  $f(x) = x^2$ . The first one is a concave function and represents a system in which fixing initial damages can return power to a higher number of customers and significantly improves utility level of service. In other words, these systems can recover from a damaging event more quickly and thus, are more resilient. However, the third utility function is convex and represents a system in which repairing the last damaged components plays a more significant role in increasing the utility level. That is, the system does not recover until higher rate of damages are fixed, which represents a less resilient system. The second utility function represents a case between the previous two systems in which fixing each damage has a constant effect in increasing the utility level. Figure 5.7 represents model's optimal solutions for these three different utility functions. Results show that in the first case, where  $f(x) = \sqrt{x}$ , the system reaches out to any utility level with lower cost compared to the other two utility functions. In other words, as we have a more resilient system, lower cost is imposed to the system to reach out a specific level of utility.

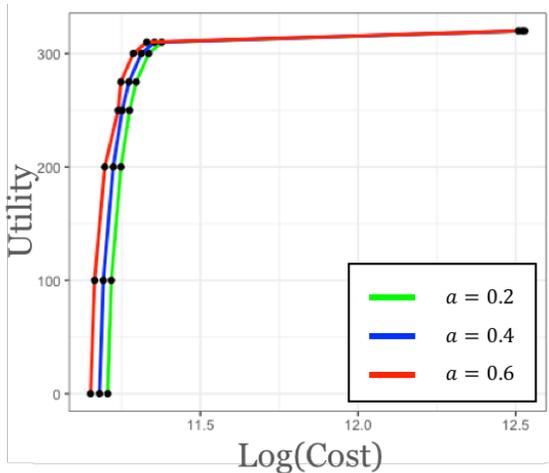


Figure 5.4: Sensitivity of the model to parameter  $a$ .

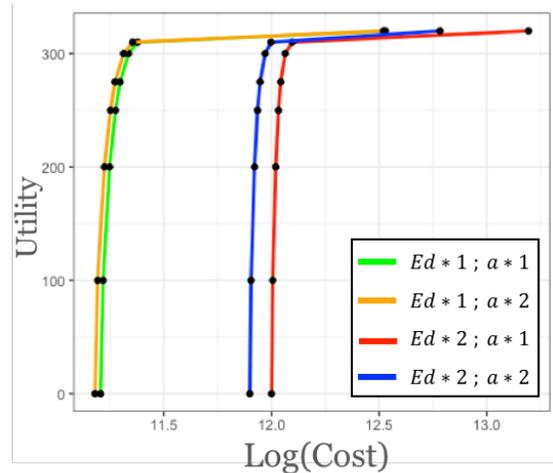


Figure 5.5: Sensitivity of the model to parameter  $Ed$ .

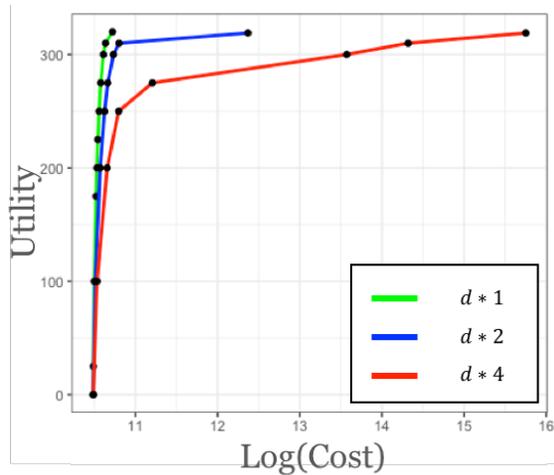


Figure 5.6: Sensitivity of the model to increase in demand.

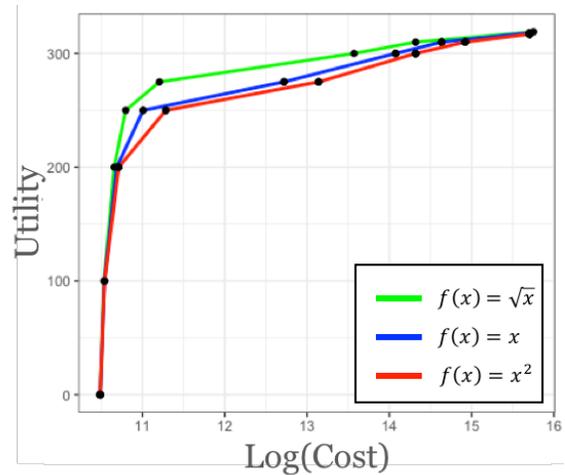


Figure 5.7: Sensitivity of the model to the utility function  $f(\cdot)$ .

## 5.6 Summary and Conclusions

In this chapter, I studied a repair crew coordination problem (RCCP) in which demands for repair crews were stochastic. Under a multi-period planning horizon, a novel multi-stage stochastic program (MSSP) was developed for the RCCP by using non-anticipativity constraints. The objective of the MSSP was to simultaneously maximize the utility level of service and minimize the costs. These two objective were combined using  $\epsilon$ -constraint method. Furthermore, a triangle method is used to convert the utility function into a linear approximation function. To deal with uncertainty in the potential crew need of districts, the LHS method was applied to generate a set of scenarios and then, using the forward scenario selection method, the number of scenarios was decreased in such a way that I achieved an appropriate scenario tree. The Gurobi solver was used for solving the problems. Using some numerical experiments, the influences of increase in crew need and the ratio of crews that can be sent to other places on the network costs and utility of service were highlighted. As this chapter introduced a novel MSSP for RCCP, there are some interesting opportunities for future researches such as considering uncertainty for other key parameters and other methods for modeling uncertainty such as robust optimization. This research was one of a few studies that applied the MSSP for RCCP for power system restoration.

## CHAPTER VI

# Conclusions and Future Research

### 6.1 Summary and Conclusions

Severe weather events have the potential to cause significant disruptions to the electric power grid. The resulting damages are, in some cases, very expensive and time-consuming to repair and they lead to substantial burdens on both utilities and customers. The frequency of such weather events has also been increasing over the past three decades and studies shows that both the number and severity of them will increase due to global warming and climate change. An important part of managing weather-induced power outages is being properly prepared for them, and this is tied in with broader goals of enhancing power system resilience. Modeling impacts of extreme weather events on the power system is a critical part of pre-storm resiliency practices because it directly influences the decisions made prior to, during, and after the event.

Accordingly, this dissertation is fundamentally motivated by the question of how we can develop implementable models to improve the resiliency of our electric power systems in the face of hazards. In particular, the dissertation has been geared towards leveraging advanced analytical tools such as data and risk analytics, statistical machine learning, and optimization for (i) improving the existing outage predictive models in terms of both accuracy and applicability, and (ii) introducing data-driven decision-making frameworks that use forecast outputs for driving better restoration and resiliency policies.

Chapter II focused on developing predictive models that can handle the zero-inflated issue in power outage data. Zero-inflation occurs whenever there exist significantly more observations of zero outages than non-zero and it is a common issue in power outage data recorded in resolutions smaller than census tract or county level. This issue leads to bias and inaccuracy in predictive modeling. Power outages are also stochastic and there always exists irreducible variability in outage predictions. The second focus of Chapter II is to develop models to accurately estimate power outages in terms of probability distributions to better

address inherent stochasticity and uncertainty in predictions. In this chapter, I proposed a novel two-stage approach integrating mixture models with resampling and cost-sensitive learning for predicting the distribution of thunderstorm-induced power outages.

The first stage is based on random forest, boosting tree and support vector machine classifiers and the second stage is based on the quantile regression forest (QRF) model. First-stage models classified data into zero class and non-zero class. In the second stage, there are two QRF models one of them trained on the zero class data and another one trained on the non-zero class data. Conditioning on the fact that each record belongs to the zero or non-zero class data, each QRF makes a separate prediction for the full distribution of that record. The role of the first-stage classifier is to predict the probability of the outcome variable being non-zero. Once this probability is estimated, a large number of random samples between 0 and 1 are generated. Then each random sample is compared with the probability of the outcome being non-zero. For each random sample larger than the estimated probability, a data point is randomly generated from the predicted distribution by the  $QRF^0$ , while for each random sample smaller than the estimated probability, a data point is randomly generated from the predicted distribution by the  $QRF^1$ . These data points together estimate the full probability distribution of each record from first stage.

The models are trained and validated using the actual thunderstorm data obtained from a decade of data collection in Alabama. The studied area is divided into grid cells and all the data and predictions are produced per grid cell. Validating my models through holdout analysis, I demonstrate that my approach offers more accurate point and probabilistic predictions compared to traditional approaches. Comparing with the traditional two-stage modeling approach, the results of holdout analysis indicate that the proposed two-stage framework improves the accuracy of the point estimates. It is also found that applying cost-sensitive learning techniques in the first-stage results in not only more precise and computationally efficient point predictions, but also higher accuracy in probabilistic predictions. More accurate predictions produced by my modeling framework help utility companies make better decisions for post-storm restoration. The probabilistic predictions help them incorporate the existing uncertainty in the predictions in their decision making process.

In chapter III, I proposed a new adaptive ensemble algorithm based on Bayesian model averaging (BMA) in order to address model uncertainty. This algorithm is built upon a number of competing base learners. The final prediction is made by averaging the prediction of these base learners where the weight of each base learner in the final prediction is proportional to its accuracy. A training set is assigned for each base learner. Each newly observed data point is added to the training set of the model making the best prediction for that record. Periodically, base learners are updated based on their new training sets. In

the proposed algorithm, unlike classical BMA approach, the weights of the base learners are based on a multinomial logistic function of the data. The weight of each base learner for a newly observed data point is different and based on the features of this record. The posterior distribution of the multinomial logit model's parameters are approximated by using the Laplace approximation method. Then, a stochastic gradient ascent approach is deployed to estimate the parameters of posterior distributions.

I validate my algorithm based on real dataset of daily customers interruptions. My case study is the first all-weather model developed for predicting customers interruptions. The results of holdout analysis show that my algorithm results in a more accurate probabilistic prediction compared to its base learners. It also provides more insights into the data generating process, and so, results in better support for utility restoration planning. Although my work is motivated by the power system application, my methodology and insights can be extended to other predictive modeling problems in which there are model uncertainty and data is collated gradually.

Due to the ease of collecting outage variables through an automated system, existing research has focused mostly on modeling the number of outages, number of customers without power, and power outage duration. However, outage focused predictive modeling is not very applicable for making system reinforcement decisions at the asset level. Inspired by this challenge, Chapter IV focused on the failure of utility assets including conductors, transformers, and poles and studied the impacts of meteorological variables on the failure of these assets. In this chapter, I developed a Bayesian belief network to model the stochastic interconnection between various meteorological factors and physical damage to different power system assets. Hypothesis tests, matching for controlling confounders effects, maximum relevant explanation, and mutual information are the tools I use to perform belief propagation and variable importance analysis. This study was based on a real dataset of daily damage occurring in 29 districts of 10 U.S. states, which are served by six operating companies. This chapter provided several critical insights that can help the policy maker (i) understand the effects of each individual variable on the power system damages, (ii) find the weather conditions that derive the maximum level of damages, and (iii) rank the meteorological factors based on their influence on the power system damages.

Finally, in Chapter V I focused on developing an optimization model that uses power outage predictions for optimally allocating repair crews to disastrous areas before and during a hazard to reduce restoration time. Based on multiple meetings with utility personnel, I developed a new multi-stage stochastic program to simultaneously make resource allocation and relocation decisions such that costs and customer satisfaction are optimized. In the proposed model, the triangle method was used to turn the nonlinear utility function to a

linear one, and  $\epsilon$ -constraint method is utilized to convert the bi-objective problem in a single objective one. I address a multi-period problem in which utility service zones (districts) have stochastic demand for repair crews. Existing uncertainty in potential demands of customer zones is modeled through a finite set of scenarios, described in the form of a scenario tree. The LHS method was applied to generate a set of scenarios and then, using the forward scenario selection method, the number of scenarios was decreased in such a way that I achieved an appropriate scenario tree. The multi-stage stochastic problem is formulated as a mixed-integer linear programming model and the Gurobi solver was used for solving the problems. Numerical results demonstrate the significance of the stochastic model. Finally, several key managerial and practical insights in terms of resource allocation are highlighted. The model developed in this study as well as the results and insights can help utility companies make better resource assignment decisions in advance of a storm.

## 6.2 Future Research

In summary, I addressed two major research areas concerning storm impacts on power systems; however, several more avenues of research, with both methodological contribution and practical impact, can be conducted to build on this thesis.

In the context of outage predictive modeling, new algorithms that capture other aspects of a resilient power system could be developed. This includes the multifaceted concept that requires an integrated approach to simultaneously predict multiple interconnected system attributes. Neural networks are recommended because on one side, they are able to model multivariate response variables and on the other side, if they are trained deep enough, are able to learn from big data and obtain the highest level of performance. Furthermore, existing deep learning regressors consider either balanced or moderately imbalanced data, and ignore the challenge of learning from significantly imbalanced data. Thus, future work is recommended for developing imbalanced deep learning frameworks for simultaneously estimating multiple highly imbalanced loss attributes for our critical infrastructures

Future work related to Chapter V could explore incorporating prioritization in recovery planning. When the extent of disaster damage to infrastructure systems is severe and widespread throughout an area, repair crews cannot respond to every damage at once. This calls for establishing prioritization of recovery activities. They are mostly prioritized as handling public safety hazards, recovering emergency facilities, repairing damages that restore service to the greatest number of customers, and moving forward to lower levels. Crew needs also are highly variable due to substantial uncertainty in damage rates. Accordingly, developing a modular and tractable data-driven decision-making framework, which (i) treats

such significant uncertainty via adaptive distributionally robust and stochastic optimization models and (ii) guarantees priority-based crew coordination planning is recommended. This allows policy makers to make optimal use of data that is revealed as time progresses and adjust their decisions based on the uncertainty realization.

Based on the evidence I have provided, I believe the model I proposed in Chapter V can be relatively easy to implement in utility companies from a technical perspective. Finally, the work presented in this thesis provides a foundation for future research on resource allocation for power system restoration.

## BIBLIOGRAPHY

- [1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [2] Ali Arab, Amin Khodaei, Suresh K Khator, Kevin Ding, Valentine A Emesih, and Zhu Han. Stochastic pre-hurricane restoration planning for electric power systems infrastructure. *IEEE Transactions on Smart Grid*, 6(2):1046–1054, 2015.
- [3] Ali Arab, Eylem Tekin, Amin Khodaei, Suresh K Khator, and Zhu Han. Dynamic maintenance scheduling for power systems incorporating hurricane effects. In *2014 IEEE International Conference on Smart Grid Communications*, pages 85–90. IEEE, 2014.
- [4] Anmar Arif, Zhaoyu Wang, Jianhui Wang, and Chen Chen. Power distribution system outage management with co-optimization of repairs, reconfiguration, and DG dispatch. *IEEE Transactions on Smart Grid*, 9(5):4109–4118, 2017.
- [5] Anil Aswani, Philip Kaminsky, Yonatan Mintz, Elena Flowers, and Yoshimi Fukuoka. Behavioral modeling in weight loss interventions. *European Journal of Operational Research*, 272(3):1058–1072, 2019.
- [6] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, Jun 2004.
- [7] Bruno SMC Borba, Marcio Z Fortes, Leonardo A Bitencourt, Vitor H Ferreira, Renan S Maciel, Marcio AR Guimaraens, Gilson BA Lima, Eduardo U Barboza, Henrique O Henriques, Nissia CR Bergiante, et al. A review on optimization methods for workforce planning in electrical distribution utilities. *Computers & Industrial Engineering*, 135:286–298, 2019.
- [8] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [9] Richard J Campbell and Sean Lowry. Weather-related power outages and electric system resiliency. Congressional Research Service, Library of Congress Washington, DC, 2012.
- [10] Luis M de Campos. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7(Oct):2149–2187, 2006.

- [11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, Jun 2002.
- [12] Chen Chen, Jianhui Wang, Feng Qiu, and Dongbo Zhao. Resilient distribution system by microgrids formation after natural disasters. *IEEE Transactions on Smart Grid*, 7(2):958–966, Mar. 2016.
- [13] Po-Chen Chen, Tatjana Dokic, and Mladen Kezunovic. The use of big data for outage management in distribution systems. In *International Conference on Electricity Distribution Workshop*, 2014.
- [14] Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1-2):43–90, 2002.
- [15] Carleton Coffrin, Pascal Van Hentenryck, and Russell Bent. Strategic stockpiling of power system supplies for disaster recovery. In *2011 IEEE Power and Energy Society General Meeting*, pages 1–8. IEEE, 2011.
- [16] Thomas M Cover, Joy A Thomas, and John Kieffer. Elements of information theory. *SIAM Review*, 36(3):509–510, 1994.
- [17] Nello Cristianini and John Shawe Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, Mar. 2000.
- [18] Denver Dash and Marek J Druzdzel. Robust independence testing for constraint-based learning of causal structure. In *UAI*, volume 3, pages 167–174, 2003.
- [19] Rachel A Davidson, Haibin Liu, Isaac K Sarpong, Peter Sparks, and David V Rosowsky. Electric power distribution system performance in Carolina hurricanes. *Natural Hazards Review*, 4(1):36–45, 2003.
- [20] Cassio P De Campos and Qiang Ji. Efficient structure learning of Bayesian networks using constraints. *Journal of Machine Learning Research*, 12:663–689, 2011.
- [21] D Dinis, AP Teixeira, and C Guedes Soares. Probabilistic approach for characterising the static risk of ships using Bayesian networks. *Reliability Engineering & System Safety*, page 107073, 2020.
- [22] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164, Aug. 1999.
- [23] Jitka Dupačová. Multistage stochastic programs: The state-of-the-art and selected bibliography. *Kybernetika*, 31(2):151–174, 1995.
- [24] Jitka Dupačová, Nicole Gröwe-Kuska, and Werner Römisch. Scenario reduction in stochastic programming. *Mathematical Programming*, 95(3):493–511, 2003.

- [25] Wei Fan, Salvatore J Stolfo, Junxin Zhang, and Philip K Chan. AdaCost: misclassification cost-sensitive boosting. In *Proceedings of 16th International Conference on Machine Learning*, volume 99, pages 97–105, Jun 1999.
- [26] Yi-Ping Fang and Enrico Zio. An adaptive robust framework for the optimization of the resilience of interdependent infrastructures under natural hazards. *European Journal of Operational Research*, 276(3):1119–1136, 2019.
- [27] Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.
- [28] David Fletcher. *Model Averaging*. Springer, 2018.
- [29] Dimitris Fouskakis. Bayesian variable selection in generalized linear models using a combination of stochastic optimization methods. *European Journal of Operational Research*, 220(2):414–422, 2012.
- [30] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Proceedings of 13th International Conference on Machine Learning*, volume 96, pages 148–156, Jul. 1996.
- [31] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, Jul. 2012.
- [32] Gina Galindo and Rajan Batta. Review of recent developments in OR/MS research in disaster operations management. *European Journal of Operational Research*, 230(2):201–211, 2013.
- [33] Vinicius J Garcia, Daniel P Bernardon, Alzenira R Abaide, Olinto A Bassi, and Guilherme Dhein. Reliability assessment by coordinating maintenance vehicles in electric power distribution systems. *Procedia-Social and Behavioral Sciences*, 111:1045–1053, 2014.
- [34] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, Mar. 2007.
- [35] Emilia Grass and Kathrin Fischer. Two-stage stochastic programming in disaster management: A literature survey. *Surveys in Operations Research and Management Science*, 21(2):85–100, 2016.
- [36] Seth D Guikema, Rachel A Davidson, and Haibin Liu. Statistical models of the effects of tree trimming on power system outages. *IEEE Transactions on Power Delivery*, 21(3):1549–1557, 2006.

- [37] Seth D Guikema and Steven M Quiring. Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data. *Reliability Engineering & System Safety*, 99:178–182, 2012.
- [38] Seth D Guikema, Steven M Quiring, and Seung-Ryong Han. Prestorm estimation of hurricane damage to electric power distribution systems. *Risk Analysis*, 30(12):1744–1752, 2010.
- [39] Seth David Guikema, Roshanak Nateghi, Steven M Quiring, Andrea Staid, Allison C Reilly, and Michael Gao. Predicting hurricane power outages to support storm response planning. *IEEE Access*, 2:1364–1373, 2014.
- [40] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, May 2017.
- [41] Seung-Ryong Han, Seth D Guikema, and Steven M Quiring. Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models. *Risk Analysis*, 29(10):1443–1453, 2009.
- [42] Seung-Ryong Han, Seth D Guikema, Steven M Quiring, Kyung-Ho Lee, David Rosowsky, and Rachel A Davidson. Estimating the spatial distribution of power outages during hurricanes in the gulf coast region. *Reliability Engineering & System Safety*, 94(2):199–210, 2009.
- [43] Seung-Ryong Han, David Rosowsky, and Seth Guikema. Integrating models and data to estimate the structural reliability of utility poles during hurricanes. *Risk Analysis*, 34(6):1079–1094, 2014.
- [44] Jichao He, David W Wanik, Brian M Hartman, Emmanouil N Anagnostou, Marina Astitha, and Maria EB Frediani. Nonparametric tree-based predictive modeling of storm outages on an electric distribution network. *Risk Analysis*, 37(3):441–458, 2017.
- [45] David Heckerman, Dan Geiger, and David M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.
- [46] Holger Heitsch and Werner Romisch. Generation of multivariate scenario trees to model stochasticity in power management. In *2005 IEEE Russia Power Tech*, pages 1–7. IEEE, 2005.
- [47] Maria Camila Hoyos, Ridley S Morales, and Raha Akhavan-Tabatabaei. Or models with stochastic components in disaster operations management: A literature survey. *Computers & Industrial Engineering*, 82:183–197, 2015.
- [48] Kai Huang and Shabbir Ahmed. The value of multistage stochastic programming in capacity planning under uncertainty. *Operations Research*, 57(4):893–904, 2009.

- [49] David-Rios Insua, Fabrizio Ruggeri, Refik Soyer, and Simon Wilson. Advances in Bayesian decision making in reliability. *European Journal of Operational Research*, 2019.
- [50] Tommi Jaakkola, David Sontag, Amir Globerson, and Marina Meila. Learning Bayesian network structure using LP relaxations. In *Proceedings of 13th International Conference on Artificial Intelligence and Statistics*, pages 358–365, 2010.
- [51] Elnaz Kabir, Seth Guikema, and Brian Kane. Statistical modeling of tree failures during storms. *Reliability Engineering & System Safety*, 177:68–79, 2018.
- [52] Elnaz Kabir, Seth Guikema, and Steven Quiring. Adaptive two-stage Bayesian model averaging for estimating the impact of hazards on power system service. *Available at SSRN, abstract id 3743815*, 2021.
- [53] Elnaz Kabir, Seth D Guikema, and Steven Quiring. Predicting thunderstorm-induced power outages to support utility restoration. *IEEE Transactions on Power Systems*, 2019.
- [54] Elnaz Kabir, Rasool Shafaei, and Esmaeil Keyvanshokoo. A novel scenario generation technique for a stochastic supply chain network. In *6th International Conference of Iranian Operations Research Society*, 2013.
- [55] Golam Kabir, Solomon Tesfamariam, Alex Francisque, and Rehan Sadiq. Evaluating risk of water mains failure using a Bayesian belief network model. *European Journal of Operational Research*, 240(1):220–234, 2015.
- [56] Peter Kall, Stein W Wallace, and Peter Kall. *Stochastic programming*. Springer, 1994.
- [57] Omar Kammouh, Paolo Gardoni, and Gian Paolo Cimellaro. Probabilistic framework to evaluate the resilience of engineering systems using Bayesian and dynamic Bayesian networks. *Reliability Engineering & System Safety*, 198:106813, 2020.
- [58] Padmavathy Kankanala, Sanjoy Das, and Anil Pahwa. Adaboost  $\hat{+}$  : An ensemble learning approach for estimating weather-related outages in distribution systems. *IEEE Transactions on Power Systems*, 29(1):359–367, 2014.
- [59] Esmaeil Keyvanshokoo, Sarah M Ryan, and Elnaz Kabir. Hybrid robust and stochastic optimization for closed-loop supply chain network design using accelerated benders decomposition. *European Journal of Operational Research*, 249(1):76–92, 2016.
- [60] Sungwoo Kim, Youngchul Shin, Gyu M Lee, and Ilkyeong Moon. Network repair crew scheduling for short-term disasters. *Applied Mathematical Modelling*, 64:510–523, 2018.
- [61] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, Nov. 2016.
- [62] Wai Lam and Fahiem Bacchus. Learning Bayesian belief networks: An approach based on the MDL principle. *Computational intelligence*, 10(3):269–293, 1994.

- [63] Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, Feb. 1992.
- [64] Edward E Leamer and Edward E Leamer. *Specification searches: Ad hoc inference with nonexperimental data*, volume 53. Wiley New York, 1978.
- [65] Jon Lee and Dan Wilson. Polyhedral methods for piecewise-linear functions i: the lambda method. *Discrete Applied Mathematics*, 108(3):269–285, 2001.
- [66] Haibin Liu, Rachel A Davidson, and Tatiyana V Apanasovich. Statistical forecasting of electric power restoration times in hurricanes and ice storms. *IEEE Transactions on Power Systems*, 22(4):2270–2279, 2007.
- [67] Haibin Liu, Rachel A Davidson, and Tatiyana V Apanasovich. Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms. *Reliability Engineering & System Safety*, 93(6):897–912, 2008.
- [68] Haibin Liu, Rachel A Davidson, David V Rosowsky, and Jerry R Stedinger. Negative binomial regression of electric power outages in hurricanes. *Journal of Infrastructure Systems*, 11(4):258–267, 2005.
- [69] David Lubkeman and Danny E Julian. Large scale storm outage management. In *IEEE Power Engineering Society General Meeting, 2004.*, pages 16–22. IEEE, 2004.
- [70] David Lubkeman, Danny E Julian, Martin Bass, and J Rafael Ochoa. Electric utility storm outage management, March 7 2006. US Patent 7,010,437.
- [71] Shanshan Ma, Liu Su, Zhaoyu Wang, Feng Qiu, and Ge Guo. Resilience enhancement of distribution grids against extreme weather events. *IEEE Transactions on Power Systems*, 33(5):4842–4853, Sep. 2018.
- [72] David Madigan and Adrian E Raftery. Model selection and accounting for model uncertainty in graphical models using Occam’s window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.
- [73] George Mavrotas. Effective implementation of the  $\varepsilon$ -constraint method in multi-objective mathematical programming problems. *Applied Mathematics and Computation*, 213(2):455–465, 2009.
- [74] D Brent McRoberts, Steven M Quiring, and Seth D Guikema. Improving hurricane power outage prediction models through the inclusion of local environmental factors. *Risk Analysis*, 38(12):2722–2737, 2018.
- [75] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, Jun 2006.
- [76] Luis Mena and Jesus A Gonzalez. Symbolic one-class learning from imbalanced datasets: application in medical diagnosis. *International Journal on Artificial Intelligence Tools*, 18(02):273–309, Apr. 2009.

- [77] Chung-ki Min and Arnold Zellner. Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics*, 56(1-2):89–118, 1993.
- [78] Mohammad Modarres. Power outage planning. *European Journal of Operational Research*, 49(2):254–265, 1990.
- [79] Nazanin Morshedlou, Andrés D González, and Kash Barker. Work crew routing problem for infrastructure network restoration. *Transportation Research Part B: Methodological*, 118:66–89, 2018.
- [80] James W Myers, Kathryn Blackmond Laskey, and Tod S Levitt. Learning Bayesian networks from incomplete data with stochastic search algorithms. *arXiv preprint arXiv:1301.6726*, 2013.
- [81] Roshanak Nateghi, Seth Guikema, and Steven M Quiring. Forecasting hurricane-induced power outage durations. *Natural Hazards*, 74(3):1795–1811, Dec. 2014.
- [82] Roshanak Nateghi, Seth Guikema, and Steven M Quiring. Power outage estimation for tropical cyclones: Improved accuracy with simpler models. *Risk Analysis*, 34(6):1069–1078, 2014.
- [83] Roshanak Nateghi, Seth D Guikema, and Steven M Quiring. Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes. *Risk Analysis*, 31(12):1897–1906, 2011.
- [84] Anders Olsson, Göran Sandberg, and Ola Dahlblom. On latin hypercube sampling for structural reliability analysis. *Structural Safety*, 25(1):47–68, 2003.
- [85] Sascha Ott, Seiya Imoto, and Satoru Miyano. Finding optimal models for small gene networks. In *Biocomputing 2004*, pages 557–567. World Scientific, 2003.
- [86] Judea Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann Publishers Inc., 1988.
- [87] Nathalie Perrier, Bruno Agard, Pierre Baptiste, Jean-Marc Frayret, André Langevin, Robert Pellerin, Diane Riopel, and Martin Trepanier. A survey of models and algorithms for emergency response logistics in electric distribution systems. part I: Reliability planning with fault considerations. *Computers & Operations Research*, 40(7):1895–1906, 2013.
- [88] Nathalie Perrier, Bruno Agard, Pierre Baptiste, Jean-Marc Frayret, André Langevin, Robert Pellerin, Diane Riopel, and Martin Trépanier. A survey of models and algorithms for emergency response logistics in electric distribution systems. part II: Contingency planning level. *Computers & Operations Research*, 40(7):1907–1922, 2013.
- [89] Feng Qiu, Jianhui Wang, Chen Chen, and Jianzhong Tong. Optimal black start resource allocation. *IEEE Transactions on Power Systems*, 31(3):2493–2494, May 2016.

- [90] Steven M Quiring, Andrea B Schumacher, and Seth D Guikema. Incorporating hurricane forecast uncertainty into a decision-support application for power outage modeling. *Bulletin of the American Meteorological Society*, 95(1):47–58, Jan. 2014.
- [91] Steven M Quiring, Laiyin Zhu, and Seth D Guikema. Importance of soil and elevation characteristics for modeling hurricane-induced power outages. *Natural Hazards*, 58(1):365–390, 2011.
- [92] Adrian E Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–164, 1995.
- [93] Adrian E Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.
- [94] Dorothy A Reed. Electric utility distribution analysis for extreme winds. *Journal of Wind Engineering and Industrial Aerodynamics*, 96(1):123–140, Jan. 2008.
- [95] Teemu Roos, Tomi Silander, Petri Kontkanen, and Petri Myllymaki. Bayesian network structure learning using factorized NML universal models. In *2008 Information Theory and Applications Workshop*, pages 272–276. IEEE, 2008.
- [96] Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.
- [97] Arif I Sarwat, Mohammadhadi Amini, Alexander Domijan, Aleksandar Damnjanovic, and Faisal Kaleem. Weather-based interruption prediction in the smart grid utilizing chronological data. *Journal of Modern Power Systems and Clean Energy*, 4(2):308–315, 2016.
- [98] Robert E Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, Jun 1990.
- [99] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [100] Marco Scutari. Learning Bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.
- [101] Glenn Shafer and Judea Pearl. *Readings in uncertain reasoning*. Morgan Kaufmann Publishers Inc., 1990.
- [102] Thomas C Sharkey, Burak Cavdaroglu, Huy Nguyen, Jonathan Holman, John E Mitchell, and William A Wallace. Interdependent network restoration: On the value of information-sharing. *European Journal of Operational Research*, 244(1):309–321, 2015.
- [103] Sarah G sharkey2015interdependent, Burak Cavdaroglu, John E Mitchell, Thomas C Sharkey, and William A Wallace. Restoring infrastructure systems: An integrated network design and scheduling (inds) problem. *European Journal of Operational Research*, 223(3):794–806, 2012.

- [104] J McLean Sloughter, Tilmann Gneiting, and Adrian E Raftery. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, 105(489):25–35, 2010.
- [105] Krishna Sridharan and Noel N Schulz. Outage management through AMR systems using an intelligent data filter. *IEEE Transactions on Power Delivery*, 16(4):669–675, 2001.
- [106] Mark FJ Steel. Model averaging and its use in economics. *arXiv:1709.08221*, 2017.
- [107] Wei Sun, Chen-Ching Liu, and Li Zhang. Optimal generator start-up strategy for bulk power system restoration. *IEEE Transactions on Power Systems*, 26(3):1357–1366, Aug. 2011.
- [108] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, Dec. 2007.
- [109] Kayu Tang, David J Parsons, and Simon Jude. Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system. *Reliability Engineering & System Safety*, 186:24–36, 2019.
- [110] David MJ Tax and Robert PW Duin. Support vector data description. *Machine Learning*, 54(1):45–66, Jan. 2004.
- [111] Kai Ming Ting. A comparative study of cost-sensitive boosting algorithms. In *Proceedings of the 17th International Conference on Machine Learning*, Jun 2000.
- [112] Gina L Tonn, Seth D Guikema, Celso M Ferreira, and Steven M Quiring. Hurricane isaac: a longitudinal analysis of storm characteristics and power outage risk. *Risk Analysis*, 36(10):1936–1947, 2016.
- [113] Hahn Tram. Technical and operation considerations in using smart metering for outage management. In *2008 IEEE/PES Transmission and Distribution Conference and Exposition*, pages 1–3. IEEE, 2008.
- [114] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.
- [115] Pascal Van Hentenryck, Carleton Coffrin, Russell Bent, et al. Vehicle routing for the last mile of power system restoration. In *Proceedings of the 17th Power Systems Computation Conference (PSCC11), Stockholm, Sweden*. Citeseer, 2011.
- [116] Chong Wang, Yunhe Hou, Feng Qiu, Shunbo Lei, and Kai Liu. Resilience enhancement with sequentially proactive operation strategies. *IEEE Transactions on Power Systems*, 32(4):2847–2857, Jul. 2017.

- [117] Shaojun Wang, Bhaba R Sarker, Lawrence Mann Jr, and Evangelos Triantaphyllou. Resource planning and a depot location model for electric power restoration. *European Journal of Operational Research*, 155(1):22–43, 2004.
- [118] Shuo Wang and Xin Yao. Diversity analysis on imbalanced data sets using ensemble models. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pages 324–331, Mar. 2009.
- [119] Yezhou Wang, Chen Chen, Jianhui Wang, and Ross Baldick. Research on resilience of power systems under natural disasters a review. *IEEE Transactions on Power Systems*, 31(2):1604–1613, Mar. 2016.
- [120] Zhaoyu Wang and Jianhui Wang. Self-healing resilient distribution systems based on sectionalization into microgrids. *IEEE Transactions on Power Systems*, 30(6):3139–3149, Nov. 2015.
- [121] DW Wanik, EN Anagnostou, BM Hartman, MEB Frediani, and M Astitha. Storm outage modeling for an electric distribution network in northeastern usa. *Natural Hazards*, 79(2):1359, 2015.
- [122] DW Wanik, JR Parent, EN Anagnostou, and BM Hartman. Using vegetation management and lidar-derived tree height data to improve outage predictions for electric utilities. *Electric Power Systems Research*, 146:236–245, 2017.
- [123] James Winkler, Leonardo Duenas-Osorio, Robert Stein, and Devika Subramanian. Performance assessment of topologically diverse power systems subjected to hurricane events. *Reliability Engineering & System Safety*, 95(4):323–336, 2010.
- [124] Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20:557–585, 1921.
- [125] Xianchao Xie and Zhi Geng. A recursive method for structural learning of directed acyclic graphs. *Journal of Machine Learning Research*, 9(Mar):459–483, 2008.
- [126] Le Xu, Mo-Yuen Chow, and Leroy S Taylor. Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-algorithm. *IEEE Transactions on Power Systems*, 22(1):164–171, Feb. 2007.
- [127] Le Xu, Mo-Yuen Chow, Jon Timmis, and Leroy S Taylor. Power distribution outage cause identification with imbalanced data using artificial immune recognition system (airs) algorithm. *IEEE Transactions on Power Systems*, 22(1):198–204, Feb. 2007.
- [128] Ningxiong Xu, Seth D Guikema, Rachel A Davidson, Linda K Nozick, Zehra Çağnan, and Kabehe Vaziri. Optimizing scheduling of post-earthquake electric power restoration tasks. *Earthquake Engineering & Structural Dynamics*, 36(2):265–284, 2007.
- [129] Ming-Jong Yao and K Jo Min. Repair-unit location models for power failures. *IEEE Transactions on Engineering Management*, 45(1):57–65, 1998.

- [130] Yiyu Yao and Bing Zhou. Two Bayesian approaches to rough sets. *European Journal of Operational Research*, 251(3):904–917, 2016.
- [131] Yuling Yao, Aki Vehtari, Daniel Simpson, Andrew Gelman, et al. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007, 2018.
- [132] Li Yijing, Guo Haixiang, Liu Xiao, Li Yanan, and Li Jinling. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, 94:88–104, Feb. 2016.
- [133] Changhe Yuan, Heejin Lim, and Tsai-Ching Lu. Most relevant explanation in Bayesian networks. *Journal of Artificial Intelligence Research*, 42:309–352, 2011.
- [134] Chengwei Zhai, Thomas Ying-jeh Chen, Anna Grace White, and Seth David Guikema. Power outage prediction for natural hazards using synthetic power distribution systems. *Reliability Engineering & System Safety*, 208:107348.
- [135] Chengsi Zhang, Xiao Liu, YP Jiang, Bo Fan, and Xiang Song. A two-stage resource allocation model for lifeline systems quick response with vulnerability analysis. *European Journal of Operational Research*, 250(3):855–864, 2016.
- [136] Yujia Zhou, Anil Pahwa, and S-S Yang. Modeling weather-related failures of overhead distribution lines. *IEEE Transactions on Power Systems*, 21(4):1683–1690, 2006.