Extracting resilience metrics and load composition from utility data

by

Nichelle'Le K. Carrington

A dissertation submitted to the graduate faculty

in towards the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Electrical Engineering

Program of Study Committee:
Zhaoyu Wang, Co-major Professor
Ian Dobson, Co-major Professor
Venkataramana Ajjarapu
Jarad Niemi
Chao Hu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2022

DEDICATION

I dedicate this dissertation to the memory of my aunt Dorothy "Dot" Young-Bigelow and godmother Mrs. Jean Draper, as well as my mother Lisa, father Eugene, sister Shēa, my brothers Xavier and Eugene II, whose, sacrifices, constant support and encouragement helped me stay motivated to complete this work.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I would like to take this opportunity to thank all the people and entities who have helped me in various aspects of my journey to complete the research and writing for this dissertation. I came to Iowa State University as a novice to the power and energy system area and found a community that nourished my hunger for knowledge. First, I want to express my gratitude to my co-advisors, Dr. Ian Dobson and Dr. Zhaoyu Wang, for their guidance, patience, and steadfast support during my research and thesis writing at Iowa State University (ISU). Their unwavering support inspired me to trust in myself and work to produce several papers and give numerous presentations. I also want to thank my committee members for their efforts: Dr. Venkataramana Ajjarapu, Dr. Jarad Niemi, and Dr. Chao Hu. Thank you, Dr. Ajjarapu, for your time and patience in helping to understand power systems and faults. Dr. Niemi, thank you for all the guidance in selecting the proper statistics courses for my minor that helped me learn how to treat and model my data. Dr. Hu, thank you for all of the feedback on my research and the literary work suggestion that helped increase my knowledge of current methods. To Dr. Anne Kimber, thank you for giving me your time, patience, and guidance. I am truly grateful for all you have done for me. My colleagues, Dr. Shanshan Ma, Dr. Anmar Arif, Dr. Kai Zhou, and my research group members Yuxuan Yuan, Zixiao Ma, Fankun Bu, Qianzhi Zhang, and Rui Cheng, thank you for always being there to support me. Dr. Cimone Wright-Hamor, thank you for navigating being a Black woman pursuing a Ph.D. in the Computer and Electrical Engineering department with me; I am grateful to have you as my partner in crime and confidant. Black ISU graduate engineer family members, Dr.Charlton Campbell, Dr. Trishelle Copeland-Johnson, Kendra Allen, Chevonne McInnis, Brianna Lawton, and Roxanna Nwachukwu, thank you for providing me a family in Iowa that I could always count on. To my writing accountability partners,Micheal Chestnut ,Bridget Perry and Schontonia Davis, thank you for every session, and then I would

like to extend my deepest gratitude to Dr. Daji Qiao, who aided me in securing the Alliance for Graduate Education & the Professoriate (AGEP) Fellowship, the Department of Electrical and Computer Engineering for the opportunity to pursue my doctoral and all the support that the faculty and staff (Sara, Vicki, Nadine, and Tony) provided me that allowed me to succeed. To Dr. Barbara Woods, Dr.Kim Wayne, Ms.Thelma Harding, and Mrs.Edna Clinton, thank you for always being there for me, encouraging me to celebrate my successes, and providing a safe zone outside of my department that helped me thrive accomplish my goal.

To my parents, Eugene and Lisa Carrington, no words can express how grateful I am to be your daughter. There was never a time that you did not support me unconditionally in my endeavors. I would not have been able to succeed in life if it had not been for all the love, prayer, and sacrifices that you both have made for my siblings and me. To my siblings, Xavier (Brooke) Carrington, Shēa (Donnie) Hammonds, and Eugene II (Shanika) Carrington, thank you for all of the love, time, and dedication that each of you gave me during this time and throughout life. To my auntie babies, Makayla, Davian, Aliyah, Xavier II, Alani, and Savannah, the world is yours to conquer, and there is nothing that you cannot achieve. I will always be here to support you whenever you need me. Thank you to my aunts and uncles for every call, card sent, and prayer you have said in my name. To my best friends, Shantera Davis and TaKedra Carey, thank you for all the reality checks, pep talks, and always standing by my side. Without the support of my family and friends, this work would not have been possible. There was never a time when someone was not in my corner. I am grateful for having each of you be a part of my life and this process.

ABSTRACT

Over the years devices such as fuse cards and smart meters have been incorporated into the electric power distribution systems infrastructure to record the status of the system. These devices record the details of outages and load. This thesis shows how this utility data can be processed to offer insights into the resilience and load composition of electric power distribution systems.

This thesis quantifies the resilience of power distribution systems using historical utility data. Resilience concerns the power system's response to stress from an external disruption such as bad weather. A resilience curve describes the accumulated outages and restores that occur in the power system as time progresses in response to the disruption. Several works have used resilience curves to model the power systems response before, during, and after a disruption. We developed a method of systematically detecting and extracting resilience curves from utility data. This allows resilience events of all sizes to be analyzed. It is common to divide idealized resilience curves into distinct time-dependent phases, such as outage and restoration phases. We defined these phases for resilience curves extracted from the utility data and calculated metrics for each phase, such as restoration duration, outage duration, recovery rate, and outage duration. The resilience curves are grouped in small, medium, and large sizes to determine the characteristics of curves with similar sizes. Resilience metrics were extracted from the curves for each group size, giving the probability distribution for each metric and its mean, standard deviation, and percentiles. The quantified uncertainties in the metrics assist utilities with giving upper bound estimates on metrics such as restoration time and customers hours impacted.

The extraction of resilience metrics from the resilience curve using phases does not address the issue of outages and restores overlapping in time, and in our data these processes substantially overlapped. Our approach provides an innovatively simple and effective way to decompose re-

silience curves into a restore process and an outage process. Metrics for each process can then be calculated. Mathematical formulas were derived to fit the data and extract the metrics, such as outage duration and restore duration, as a function of the number of outages. The variability of duration resilience metrics was calculated from the decomposed processes. For each number of outages, the mean duration and standard deviation determined a gamma distribution and the upper bound of a 95% confidence interval was calculated. A function from that fitting was derived to estimate an 95% upper bound of the duration based on the number of outages. Similarly, we were able to extract the restore and outage processes from resilience curves and derived mathematical formulas for customer hours lost and risk as a function of the number of outages. These new approaches to deriving and analyzing metrics are a novel statistical analysis that works with practical utility data, avoids the complexities of modeling an individual repair process, and applies decomposition to solve the problems of overlapping processes.

The thesis also processed some transmission system utility data, showing how to obtain useful detailed records from a public website, and examining the weather impact on cascading outages.

We also developed software that processes and analyzes advance metering infrastructure (AMI) data for small utilities. AMI data is a recording of energy consumption for distribution customers at the building level and can record the energy consumption as finely as per minute. A deployable tool was developed to aid small utilities with processing AMI data. One analysis in the tool is the capability of classifying customers based on consumption. This analysis uses k-means clustering to group the customers based on load. A comprehensive breakdown of the load consumption is another analysis feature within the tool. The hourly load consumption is broken down into the contributions of each customer class. The tool was developed to provide small utilities with the capability to clean, analyze and export their AMI data.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Research Motivation and Problem Statement

Power outages disrupt daily tasks for utility customers, including businesses and the public, resulting in income and productivity losses. Ideally, the power system should be able to self-recover during a disruption by rapidly detecting faults, minimizing impact, and quickly returning to normal. A number of factors contribute to outages, including aging equipment, animals, and harsh weather, resulting in higher resilience expenses for power system utilities [1–4].

Over the years, resilience and reliability research has increased to capture the self-healing capacity of the power systems. Utilities use reliability metrics to measure the power system's reliability. Many utility metrics are available: the System Average Interruption Frequency Index (SAIDI), Consumer Average Interruption Duration Index (CAIDI), and System Average Interruption Frequency Index (SAIFI), and researchers are focusing on EENS (Expected Energy Not Served) and LOLE (Loss of Load Expectation) [5]. All of these are utilized to evaluate the system's average performance over the year [5, 6]. It is worth noting that resilience metrics are different from reliability metrics. While reliability metrics examine outages and restores throughout the year, resilience metrics consider only how the system responds during a resilience event. As far as customer impact is concerned, reliability metrics often omit extreme events, whereas resilience metrics describe how these grids respond and recover during times of severe events [7, 8].

While reliability metrics are oriented toward high-probability and low-impact events, resilience metrics are oriented towards low-probability and high-impact events [5, 8]. Indeed the literature examining real data focuses only on the most extreme events. In doing so, they neglect the contribution of information from lower probability but still impactful resilience events. Considering only the extreme cases also limits the number of observed cased. The use of reliability metrics in

combination with new resilience metrics can improve the description of both resilience and reliability of the system [9, 10].

All utilities strive to supply their customers with reliable and steady power, but they prioritize restoring power safely and efficiently when outages occur. Restoration is expensive and complex because of several variables, including scheduling, interdependencies between areas, and spatial considerations. Utility companies are unable to accurately predict outage event duration because of complexity, as was discussed in [11, 12]. The ability to provide resilience metrics such as restoration times with a degree of confidence will allow utilities to better serve their customers and prepare for outages [13].

Quantifying system resilience is a popular topic of study between many fields [13–18]. In energy, economy, telecommunications, and cyber-security infrastructure, simulation and modeling have been used to identify weaknesses [6, 14] and improve system performance under extreme conditions [19–21]. Many studies have used resilience curves as a credible method to model power systems' responses to an event and to quantify resilience in the power distribution system [10, 11, 16, 17, 19, 22, 23]. Researchers proposed resilience metrics for isolated extreme weather events, such as hurricanes, in [11, 13–17, 23–25]. Only extreme examples of resilience were examined in these studies, omitting the occurrence of minor and medium level resilience events that can lend insight into a system's overall resilience. Current methods to extract metrics (such as the restoration time and outage propagation) divide a resilience curve into distinct outage and restore phases but fail to consider the overlap between outage and restore phases that occurs in practice [6, 18, 26, 27]. In many models, the uncertainty of a resilience metric is not accounted for and the metrics are often designed without the input of real utility data. To better aid utilities in providing quality service to customers, an automatic processing method is needed for analyzing the resilience of a system with a degree of certainty.

1.2 Background and Literature Survey

1.3 Detailed Outage Data

1.3.1 Distribution Outage Data

The distribution data used in this work is from a regulated distribution utility that serves approximately 300,000 electric customers in a defined service territory that includes both urban and rural demographics. The utility gathered the outage data for six years from 2011 to 2016. The start and end time of the outages were recorded by fuse cards equipped to record the time at which an outage begins and ends based on the loss of power. The spread of the data locations covers a total area of approximately $4,800$ mi^2. The raw data is private.

1.4 Resilience, Curves and Metrics

1.4.1 Resilience in Power Systems

In power systems, resilience is defined as the ability to prepare for, absorb, adapt to, and/or rapidly recover from adverse events [7, 8, 14, 19, 22, 28]. The primary goal of an electric utility is to maintain an equilibrium at which a stable and steady supply of electricity is provided to customers. In order to provide a steady and stable electricity supply, utilities aim to ensure that their system has both operational resiliency and infrastructure resiliency. Operational resilience in a system aids a power system to maintain operational strength and robustness in the face of an adverse event [5, 6, 18]. Whereas infrastructure resilience address the physical strength of a power system to minimize impact of the portion of the system that is damaged, collapsed, impaired or nonfunctional [6, 18]. In comparison, operational infrastructure focuses on the performance of the system for operational decisions and the infrastructure resilience focuses on the physical capabilities of the system. In order to observe the infrastructure and operational resiliency of the system during an adverse event, resilience curves are used.

1.4.2 Resilience Curves

Resilience curves model a system's response before, during, and after a hazard has occurred [6, 13, 17–19, 26–29]. The resilience curve tracks the accumulated outages and restores as time progresses during an event. In order for a resilience curve to be considered by us an outage event in distribution systems, the accumulation of a group of successive outages and restores and the total number of incidents must be greater than or equal to 2. This means that the event must have at least 2 or more outages and 2 or more restores within its time period.

1.4.3 Resilience Phases

1.4.3.1 Five Stage Framework

The ideal resilience trapezoid in Figure 1.1 is a linear approximation of the resilience curve as a function of time and is used to quantify the resilience level in a multi-phase resilience framework [6]. The resilience trapezoid can be divided into five phases that characterize the progression of a power system under the influence of an external disturbance. The five phases are the pre-disturbance steady-state phase, disruptive phase, post-disruptive phase, recovery phase and the new steady state phase [18, 30]. The diagram in Figure 1.1 shows the order of each phase as an event progresses over time using the ideal resilience trapezoid. The pre-disturbance phase ($t_0 \leq t \leq t_H$) captures the system's performance prior to the hazard. The damage propagation phase ($t_H \leq t \leq t_P$) captures the system performance after a hazard has occurred and while the accumulation of successive outages and restores are in the process of degrading the system's performance state. The degraded state of the system is the third phase while the performance of the system is reduced significantly from the downed components. In this phase the assessment of the damage by inspection crews occurs, as well as the removal of any debris that is surrounding the damaged components that hinder the repair crews from accessing the area. The length of the degraded state varies from event to event due to several factors such as the availability of inspection and removal crews and the time needed for clearing the debris around components. The fourth phase is the restoration stage ($t_R \leq t \leq t_{SS}$); in this stage the process of restoring the

downed and damaged components takes place. The final and fifth stage is the new steady state phase ($t_{SS} \leq t$). In the final stage all damaged and downed lines and component have been repaired and the system is back to a steady state of functioning. The five phases of the resilience curve are depicted in Figure 1.1 over an ideal resilience trapezoid.

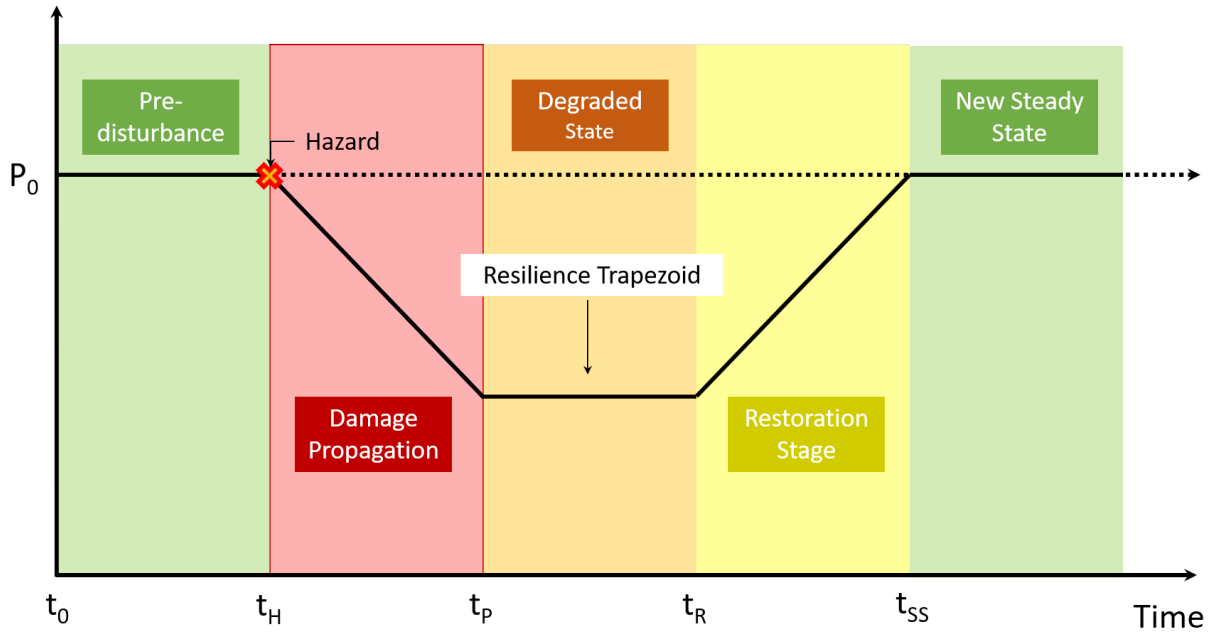

Figure 1.1: The five phases of a resilience curve shown on an ideal resilience trapezoid.

1.4.3.2 Three Stage Framework

The three-stage framework depicted on the resilience curve in Figure 1.2 is a generalized model to capture a system's recovery process as it progresses in each stage [6, 18, 19]. This simplified framework ignores the pre-disturbance and post repair new steady state phases of the five phase framework in Figure 1.1.

The first stage focuses on hazard prevention ($0 \leq t \leq t_S$). This is when the system maintains normality before a hazardous event starts. The second stage is the outage propagation ($t_S \leq t \leq t_N$), where the successive outages occur at a faster rate while the hazard is absorbed by the system. The third stage of restoration ($t_N \leq t \leq t_E$) is the recovery of the system back to

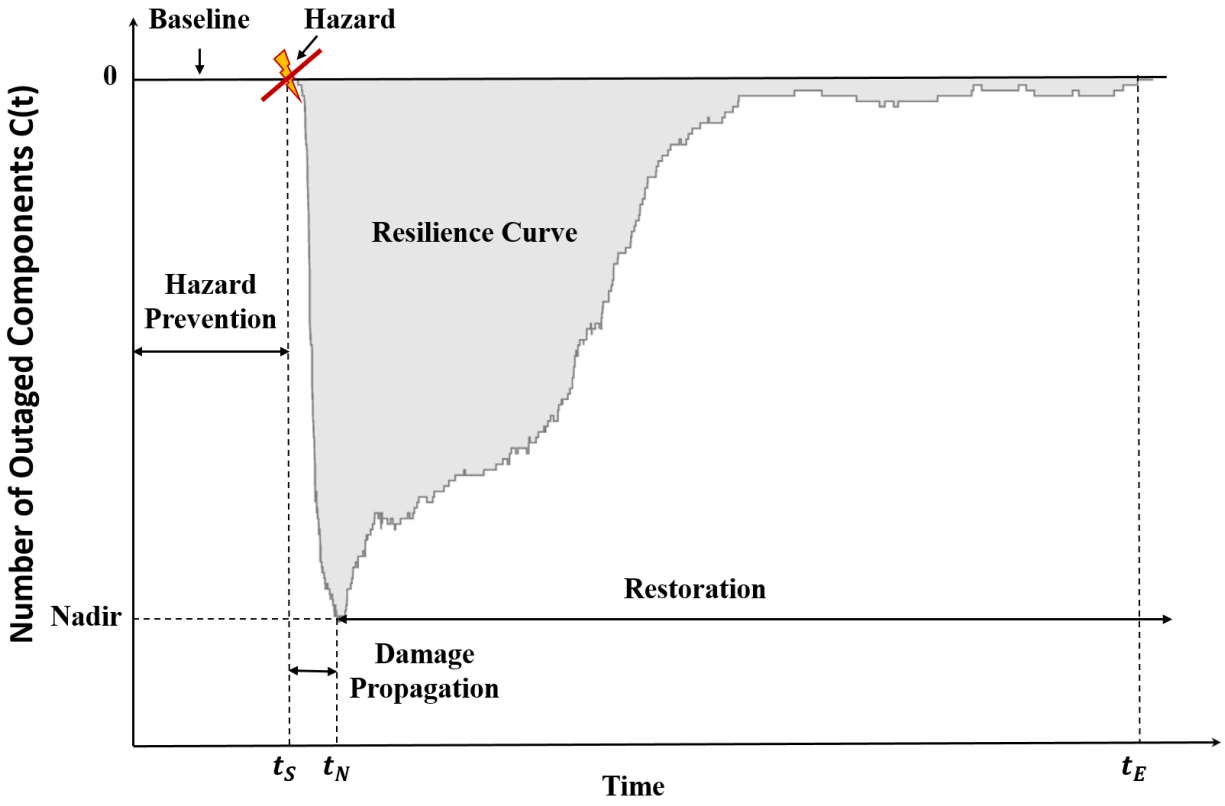

Figure 1.2: The three stages of resilience shown on an extracted resilience curve from real utility data.

normal operation. This framework is used to simplify the characterization of the resilience curve in order to quantify the resilience of a system during an event into useable metrics. The advantage of dividing the resilience curve into phases is that the portions of each phase can be used to describe aspects of the resilience of the system.

1.4.4 Resilience Metrics

Resilience curves are used in the development of resilience metrics to quantify the performance of the power system before, during and after a disruptive event. The quantification of resilience into metrics arose as a means to capture the real performance of the power system versus the ideal performance level. One way the resilience of a system can be measured is by finding the area under a resilience curve. For example, the difference in the area between the real and ideal

resilience curves and the baseline can be interpreted as a ratio of the proportion of delivery function that has been recovered from its disrupted state [6]. Although there is no standard definition of resilience in power systems there are several works that define and model resilience metrics using the dimensions of resilience curves, trapezoid and triangles as a way to capture and characterize the dynamics and performance of a system[18, 26–28, 30, 31]. For example, the metrics for the restoration phase such as restoration duration or restoration rate and derived from the resilience triangle associated with the restoration resilience. The concept of resilience metrics derived from resilience curves has been applied in numerous works to model both transmission system and distribution system outage processes.

1.5 Research Objective

This thesis presents new methods to process real utility data to improve our understanding of power system resilience. Its objectives are as follows:

- Develop a process for automatically detecting and extracting resilience events in real utility outage data. The first step uses an event definition that identifies the start and end of an outage event to automatically extract the events. Secondly, the process categorizes events according to event size. After that, it calculates resilience metrics based on a range of sizes.
- Establish a method to decompose resilience curves into decoupled outage and restore processes. Derive resilience metrics from the outage and restore processes and use the data to develop formulae that describe how the metrics depend on event size.
- Estimate the variability of resilience metrics such as restore duration to quantify uncertainty for prediction.
- Derive a formula to estimate the risk as the product of customer hours and probability and how it depends on the event size.
- Determine if the distribution system outages of different sizes have dominant cause codes. Using daily climate data, determine the impact of average wind speed on resilience events.

- Observe the effect of weather and other influences via cause codes from historical transmission utility data. Process detailed utility outage data and NOAA storm data to match climate information to outage's cause. Get bulk statistics on annual cascading showing the dependence on storms.

- Provide a new processing for outage transmission system data from a public website. Use the processed data to get bulk statistics on the automatic transmission line outages identified from the processing to quantify the resilience events propagation and spread during cascading as an example of the value of the processed data.

- Develop a research-grade, Excel-based software tool that small public utilities can extract useful information from advance metering infrastructure (AMI) data. The tool should import, process, analyze, and export large volumes of AMI data recorded at intervals of 15-minute and hourly rates.

1.6 Organization of the Dissertation

The rest of this dissertation is organized as follows. Chapter 2 addresses the problem of detecting and extracting outage events as resilience curves within utility data. The chapter starts with a definition of indicators of the start and end outage events to extract from utility data, followed by a description of how resilience curves are categorized based on the number of components outaged, and finally, provides simple statistics on each category of curves. Chapter 3 presents a model that decomposes resilience curves into an outage process and a restore process to address the problem of overlapping outages and restores. The chapter begins with the definition of mathematical formulations of resilience metrics such as restore duration and then computes statistics based on these formulations. To estimate the risk, we multiply customer hours by probability, based on the number of outages, and get the distribution of the risk with respect to event size. Chapter 4 examines distribution system outage cause codes to show the cause of outage and resilience events. Chapter 5 compresses historical utility outage data and NOAA storm

data to match climate information to outages recorded as storms by utility cause codes. The chapter observes the impact of weather and other influences on utility outages and cascades of outages. Chapter 6 presents a research-grade, standalone tool that processes and analyzes smart meter data. The chapter first introduces a data management plan, then describes the process in which the tool follows to clean, analyze data and produce graphics. Chapter 7 shows how to process bulk statistics on transmission system outage data from a public website. Then the chapter compares the resilience event propagation and spread during cascading using bulk statistics on automatic transmission line outages. Chapter 8 summarizes contributions of this work, the publications and presentations given, and outlines future work.

1.7 References

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

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

[3] H. C. Caswell, V. J. Forte, J. C. Fraser, A. Pahwa, T. Short, M. Thatcher, and V. G. Werner, "Weather normalization of reliability indices," *IEEE Transactions on Power Delivery*, vol. 26, no. 2, pp. 1273–1279, 2011.

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

[5] A. A. Bajwa, H. Mokhlis, S. Mekhlief, M. Mubin, M. M. Azam, and S. Sarwar, "Resilience-oriented service restoration modelling interdependent critical loads in distribution systems with integrated distributed generators," *IET Generation, Transmission & Distribution*, 2021.

[6] M. Panteli, P. Mancarella, D. N. Trakas, E. Kyriakides, and N. D. Hatziargyriou, "Metrics and quantification of operational and infrastructure resilience in power systems," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4732–4742, 2017.

[7] A. Pepiciello, A. Vaccaro, and L. L. Lai, "An interval mathematic-based methodology for reliable resilience analysis of power systems in the presence of data uncertainties," *Energies*, vol. 13, no. 24, 2020. [Online]. Available: https://www.mdpi.com/1996-1073/13/24/6632

[8] H. Raoufi, V. Vahidinasab, and K. Mehran, "Power systems resilience metrics: A comprehensive review of challenges and outlook," *Sustainability*, vol. 12, no. 22, 2020. [Online]. Available: https://www.mdpi.com/2071-1050/12/22/9698

[9] M. Panteli and P. Mancarella, "The grid: Stronger, bigger, smarter?: Presenting a conceptual framework of power system resilience," *IEEE Power and Energy Magazine*, vol. 13, no. 3, pp. 58–66, 2015.

[10] M. Panteli and P. Mancarella, "Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies," *Electric Power Systems Research*, vol. 127, pp. 259–270, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037877961500187X

[11] Y. Wei, C. Ji, F. Galvan, S. Couvillon, G. Orellana, and J. Momoh, "Non-stationary random process for large-scale failure and recovery of power distribution," *Applied Mathematics*, 2016.

[12] G. P. Cimellaro, *Urban Resilience for Emergency Response and Recovery: Fundamental Concepts and Applications*, 1st ed. Cham, Switzerland: Springer International Publishing, 2016.

[13] B. M. Ayyub, "Systems resilience for multihazard environments: Definition, metrics, and valuation for decision making," *Risk Analysis*, vol. 34, no. 2, pp. 340–355, 2014.

[14] B. M. Ayyub, "Practical resilience metrics for planning, design, and decision making," *ASCE-ASME Journal Risk & Uncertainty Engineering Systems, Part A: Civil Engineering*, vol. 1, no. 3, p. 04015008, 2015.

[15] C. Ji, Y. Wei, and H. V. Poor, "Resilience of energy infrastructure and services: Modeling, data analytics, and metrics," *Proceedings of the IEEE*, vol. 105, no. 7, pp. 1354–1366, 2017.

[16] D. Henry and J. E. Ramirez-Marquez, "Generic metrics and quantitative approaches for system resilience as a function of time," *Reliability Engineering & System Safety*, vol. 99, pp. 114–122, 2012.

[17] R. Francis and B. Bekera, "A metric and frameworks for resilience analysis of engineered and infrastructure systems," *Reliability Engineering & System Safety*, vol. 121, pp. 90–103, 2014.

[18] M. Panteli, D. N. Trakas, P. Mancarella, and N. D. Hatziargyriou, "Power systems resilience assessment: hardening and smart operational enhancement strategies," *Proceedings IEEE*, vol. 105, no. 7, pp. 1202–1213, 2017.

[19] M. Ouyang, L. Dueñas-Osorio, and X. Min, "A three-stage resilience analysis framework for urban infrastructure systems," *Structural safety*, vol. 36, pp. 23–31, 2012.

[20] M. Shinozuka, X. Dong, T. Chen, and X. Jin, "Seismic performance of electric transmission network under component failures," *Earthquake Engineering & Structural Dynamics*, vol. 36, no. 2, pp. 227–244, 2007.

[21] L. Dueñas-Osorio and S. M. Vemuru, "Cascading failures in complex infrastructure systems," *Structural safety*, vol. 31, no. 2, pp. 157–167, 2009.

[22] S. Ma, N. Carrington, A. Arif, and Z. Wang, "Resilience assessment of a self-healing distribution systems under extreme weather events," in *IEEE PES General Meeting*. IEEE, 2019.

[23] C. Ji, Y. Wei, H. Mei, J. Calzada, M. Carey, S. Church, T. Hayes, B. Nugent, G. Stella, M. Wallace *et al.*, "Large-scale data analysis of power grid resilience across multiple us service regions," *Nature Energy*, vol. 1, no. 5, pp. 1–8, April 2016.

[24] Y. Wei, C. Ji, F. Galvan, S. Couvillon, and G. Orellana, "Dynamic modeling and resilience for power distribution," in *2013 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2013, pp. 85–90.

[25] Y. Wei, C. Ji, F. Galvan, S. Couvillon, G. Orellana, and J. Momoh, "Non-stationary random process for large-scale failure and recovery of power distributions," *Applied Mathematics*, 2012.

[26] N. Bhusal, M. Abdelmalak, M. Kamruzzaman, and M. Benidris, "Power system resilience: Current practices, challenges, and future directions," *IEEE Access*, vol. 8, pp. 18 064–18 086, 2020.

[27] Z. Bie, Y. Lin, G. Li, and F. Li, "Battling the extreme: A study on the power system resilience," *Proceedings of the IEEE*, vol. 105, no. 7, pp. 1253–1266, 2017.

[28] N. Yodo and P. Wang, "Resilience modeling and quantification for engineered systems using bayesian networks," *Journal of Mechanical Design*, vol. 138, no. 3, p. 031404, 2016.

[29] S. Hosseini, K. Barker, and J. E. Ramirez-Marquez, "A review of definitions and measures of system resilience," *Reliability Engineering & System Safety*, vol. 145, pp. 47–61, 2016.

[30] C. Nan and G. Sansavini, "A quantitative method for assessing resilience of interdependent infrastructures," *Reliability Engineering & System Safety*, vol. 157, pp. 35–53, 2017.

[31] M. Panteli and P. Mancarella, "Modeling and evaluating the resilience of critical electrical power infrastructure to extreme weather events," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1733–1742, Sept 2017.

CHAPTER 2. EXTRACTING RESILIENCE STATISTICS FROM UTILITY DATA IN DISTRIBUTION GRIDS

Nichelle'Le K. Carrington, Shanshan Ma, Ian Dobson, and Zhaoyu Wang, Department of Electrical and Computer Engineering Iowa State University, Ames, Iowa, USA

Modified from a manuscript published in *2019 IEEE Power & Energy Society General Meeting* [1]

2.1 Abstract

It is useful to quantify electrical distribution system resilience based on historical performance. This paper systematically extracts resilience curves from historical utility outage data, extracts resilience metrics such as duration, average recovery rates, and maximum number of simultaneously outaged components, and examines the statistics of these resilience metrics for small, medium, and large events. The resilience metrics and their typical variabilities are expected to be helpful in predicting and bounding the likely outcomes of future resilience events. For example, we can calculate the restoration time that will be achieved with 95% confidence.

2.2 Overview

Maintaining a continuous energy supply to customers is the goal of utilities, but there is always threat of unplanned disruption of electrical services in power distribution systems [2]. From customers seeking information on when the power will supply will return and the several variables that impact the complex and intricate process of restoring power, utilities are under a lot of pressure to restore power supply effectively, safely, and efficiently. The current method of estimating when power can be restored is considered a "best guess" estimate of the time, due to several factors such as availability of inspection crews, repair crews, clean up time and transitional times

for crews [2, 3]. These factors typically cause utilities to have uncertainty in their the approximation of restoration time. The work in this chapter presents a method that enhances the ability to approximate restoration times by processing detailed, historical data from previous events.

Detailed outage data is routinely collected by many utilities to observe the dynamic performance of their system. Researchers use this data to aid utilities with providing quicker estimation of the outage duration and restoration time. Many researchers have used resilience curves as a credible method to model and evaluate system vulnerability and the ability to recover from hazards or adverse events. In [4], the entire life cycle of failure and recovery of large scale power failures is considered, but lesser events are ignored, which tends to exaggerate the typical impacts of events. The work in [5] statistically analyzed factors that affect outage duration but did not predict the duration or its variability. Authors in [2] used the text from inspection reports (without considering the number of outages) to predict outage duration to facilitate customer preparation. In these studies, extreme weather events are considered as isolated events and are used to observe the impacts of extreme conditions on the system's performance. But they omit the less extreme and more typical events that also contribute to the system's overall resilience.

The focus of this chapter is to develop a method to systematically detect and extract resilience curves from detailed historical outage data from a distribution utility within the United States. The resilience curves represent all the events that have disrupted the normal condition of the distribution system for this utility. The resilience events are detected by processing the cumulative number of outages as a function of time. The cumulative number of outages is the number of outage incidents present at a given time that have not yet been restored. Threshold values on the cumulative number of outages are used to define the beginning and end of the resilience events and classify the events into small, medium and large. The classification into small, medium and large events allows the resilience metrics of each size of event to be calculated, as well as the variability of the metric.

The classification of resilience curves allows the curves to be grouped into a large set and systematically analyzed instead of re-sampling from one isolated event repetitively. The outage prop-

agation and restoration stages are evaluated using resilience triangles on a large data set to assess: (a) variability of duration for outage and recovery processes to help with predictions; (b) average outage and recovery rates during events to help with assessment and predictions. By systematically detecting resilience curves, we are able to gain better insight on the overall performance of the system and generate statistics of the duration of outage and recovery processes from multiple events instead of focusing on one single event.

The rest of the chapter is organized as follows: Section 2.3 describes the event extraction procedure (2.3.1), and the metrics calculated (2.3.2). Details on the detection of customer resilience curves are presented in section 2.3.3. The metrics and their variabilities on the distribution utility data sets are presented in section 2.4, and section 2.5 concludes the chapter.

2.3 Resilience and Systematic Detection

2.3.1 Extracting Events

The cumulative number of unrestored outages varies with time as outages occur and are restored. Under normal conditions the cumulative number of unrestored outages stays near zero because outages are generally infrequent and are restored quickly. But under stressed conditions, outages are more frequent and accumulate before they can be restored, and the cumulative number of outages has excursions away from zero. These accumulations of outages are the resilience events. The resilience events are extracted from the data by detecting when the cumulative number of outages passes and returns to a threshold number of outages. We now give more details of this extraction.

Since the utility data includes the outage and restore time for each component, it is straightforward to sort these times by their order of occurrence and then calculate

$$\begin{aligned} C(t) = -(&\text{cumulative total outages at time } t \textit{ minus} \\ &\text{cumulative total restores at time } t) \\ = -(&\text{number of simultaneous outages at time } t) \end{aligned} \tag{2.1}$$

The threshold number of outages is zero or a small negative number of outages C_{base}; we use $C_{\text{base}} = 0$ as a simple case. Under normal conditions $C(t)$ is at or above C_{base}. The start time t_S of an event is defined by $C(t)$ decreasing below C_{base} and the end time t_E of an event is defined by $C(t)$ increasing to C_{base}. Then $C(t)$ for $t_S \leq t \leq t_E$ is a resilience curve for the event as shown in Fig. 1.2. The minus sign in (2.1) ensures compatibility with standard resilience curves.

2.3.2 Nadir, Resilience Triangles, and Metrics

In Figure 2.1, the lowest point of the resilience event curve $C(t)$, called the nadir, occurs at t_N. The nadir $C(t_N)$ corresponds to the maximum number of simultaneously occurring outages in an event. In the exceptional case of several low points occurring at exactly the same level in the same event, we choose the last one to be the nadir. Resilience curves can be divided into the propagation process and the restoration process by applying the three-stage framework detailed in Section 1.4.3.2. We use the nadir to locate the end of the propagation process and the beginning of the restoration process. Note that dividing the event time into separate propagation and restoration processes in this way is idealized, since in real data these processes overlap somewhat as shown in Figure 2.2. This problem will be addressed and solved in Chapter 3. Using the nadir to define phases, metrics of duration are easily obtained with the application of resilience triangles by measuring the width of the triangle. In Figure 1.2, the duration of propagation is $t_N - t_S$, and the duration of restoration is $t_E - t_N$. The event duration is $t_E - t_S$. These duration metrics can aid in explaining the impact of a disruptive event and predict the impacts of future events [6]. Average rate metrics are easily obtained from the slopes of the resilience triangles. The average outage process rate is $-C(t_N)/(t_N - t_S)$ and the average recovery process rate is $-C(t_N)/(t_E - t_N)$ in Figure 1.2.

The maximum number of simultaneously outaged components, the duration of an event, the restoration time, and the average outage and recovery rates are important metrics that we extract from previous events to help a utility estimate these metrics for an anticipated or ongoing event. In particular, we calculate the statistics of these metrics from previous events using de-

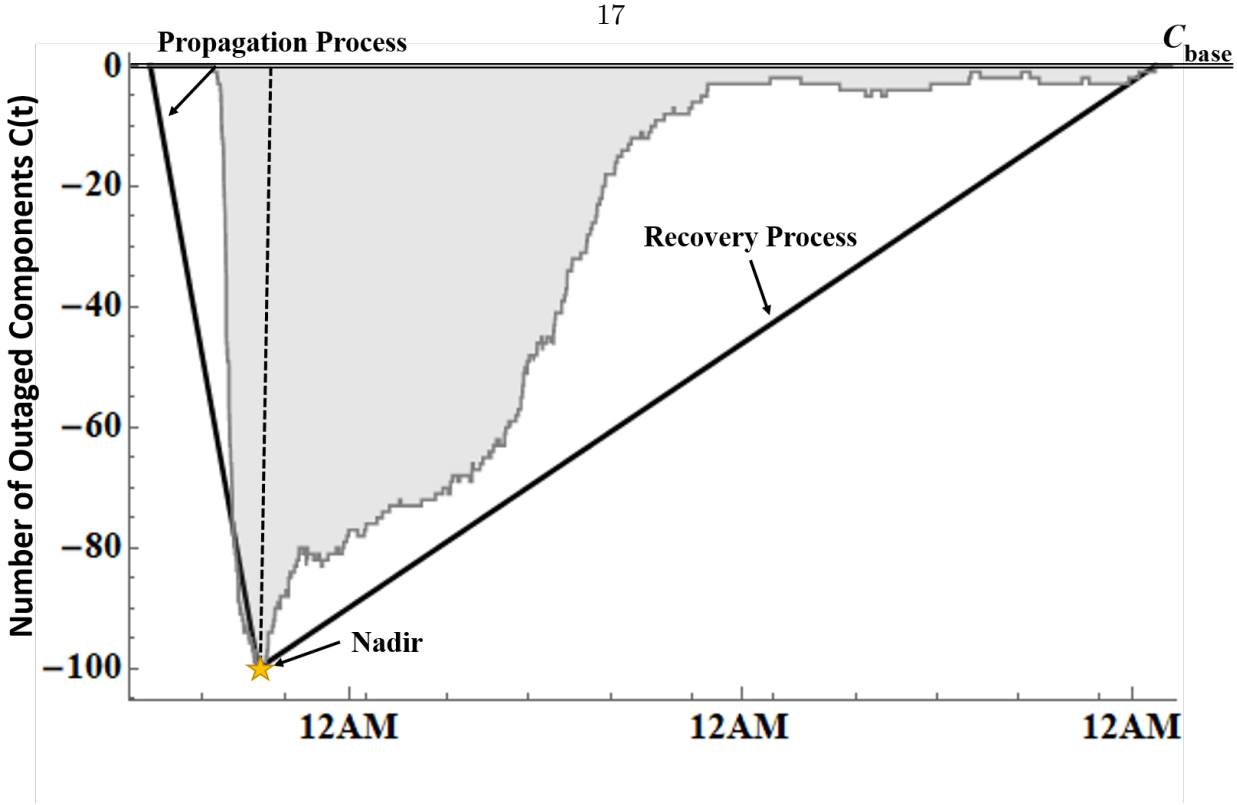

Figure 2.1: The outage propagation process begins when the curve decreases from the baseline at zero and ends at the nadir. The recovery process starts at the nadir and ends when the curve increases to the baseline.

tailed outage data to be able to estimate a 95% upper bound confidence interval for the restoration time. This can help provide the customers of the utility with an upper bound estimate of the restoration time with a reasonable certainty. Events are grouped into small, medium, or large depending on the value of the nadir $C(t_N)$ of their resilience curve:

Small events have $-3 \geq C(t_N) \geq -9$.

Medium events have $-10 \geq C(t_N) \geq -19$.

Large events have $-20 \geq C(t_N)$.

2.3.3 Customer Resilience Curves

Since the outage data also includes the number of customers outaged and restored, we can also form the cumulative number of customers out $C^{\text{cust}}(t)$ as a function of time, similarly to the definition of $C(t)$ in (2.1) except that "outages" are replaced by "customers". Then the customer

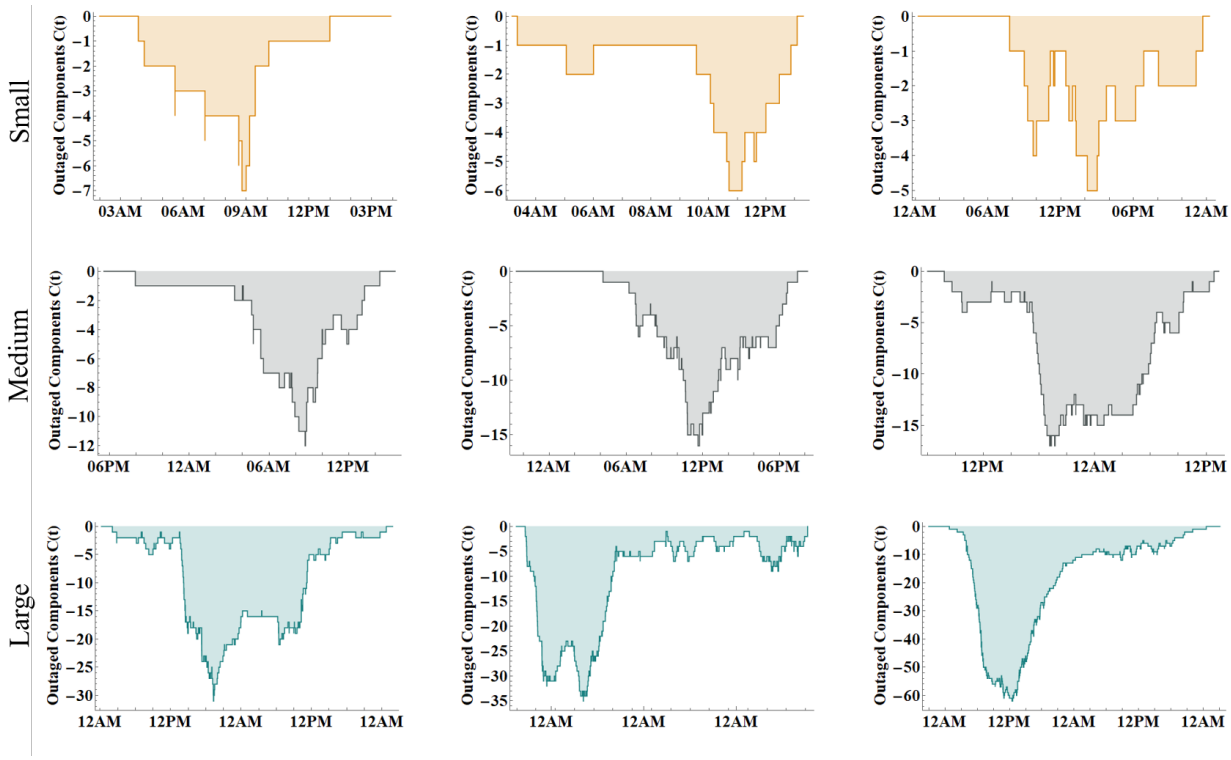

Figure 2.2: Component resilience curve examples. Small event (Orange) causes are scheduled maintenance or minor physical damage. Medium event (Gray) causes are moderate weather/storm or moderate physical damage. Large event (Teal) causes are extreme or severe weather/storm or severe physical damage.

resilience curves for an event occurring for $t_S \leq t \leq t_E$ is the portion of the cumulative customer curve $C^{\text{cust}}(t)$ for $t_S \leq t \leq t_E$ (see Fig. 2.3). The area above the customer resilience curve is the total customer hours outaged in the event. If the event were to be included in the SAIFI calculation, this customer area would directly add to the SAIFI numerator. The average customer recovery rate is $-C^{\text{cust}}(t_N)/(t_E - t_N)$.

2.4 Results

The data used for the results in this chapter is defined in section 1.3.1. The probability density functions for propagation, restoration and total event durations for all events are shown in Fig. 2.4.A breakdown of the durations by metric and group will be detailed in the rest of this section.

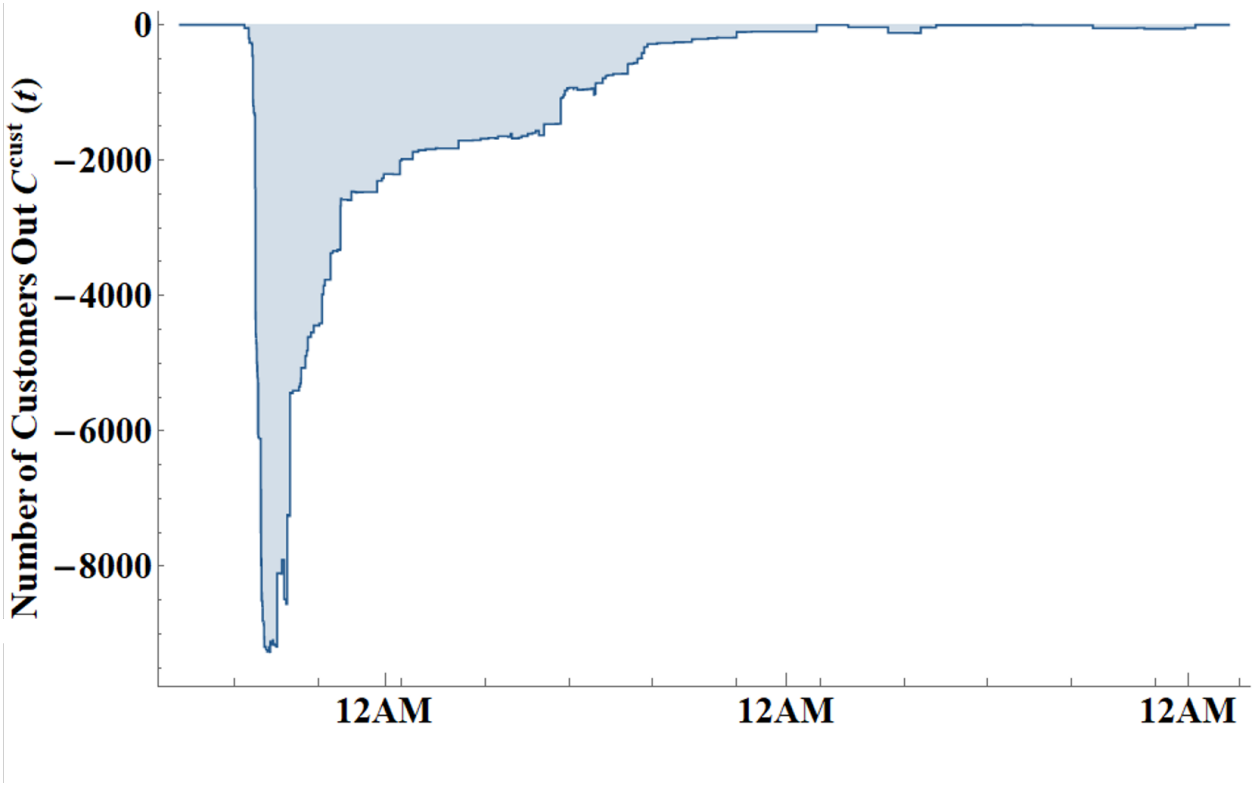

Figure 2.3: The customer resilience curve shows the cumulative number of customers outaged during the event corresponding to Figure 2.1.

2.4.1 Cause Codes

The cause code of each outage and restore within an event was tracked to determine the most common cause of outages for the event. There were 34,945 outages with causes reported and 63 types of cause codes. About 5% of those causes were weather-related, and 14% were animal-related. The remaining causes were mostly component malfunctions, tree limbs, and debris. The top causes for all events were tree limbs near the clearance zone of lines and squirrels. The top three weather-related causes were wind, rain, and lightning.

2.4.2 Small, Medium and Large Events

1486 events were extracted and grouped by size using the methods of Section 2.3. There were 910 small events, 75 medium events, and 50 large events found in the data. The events in Fig.

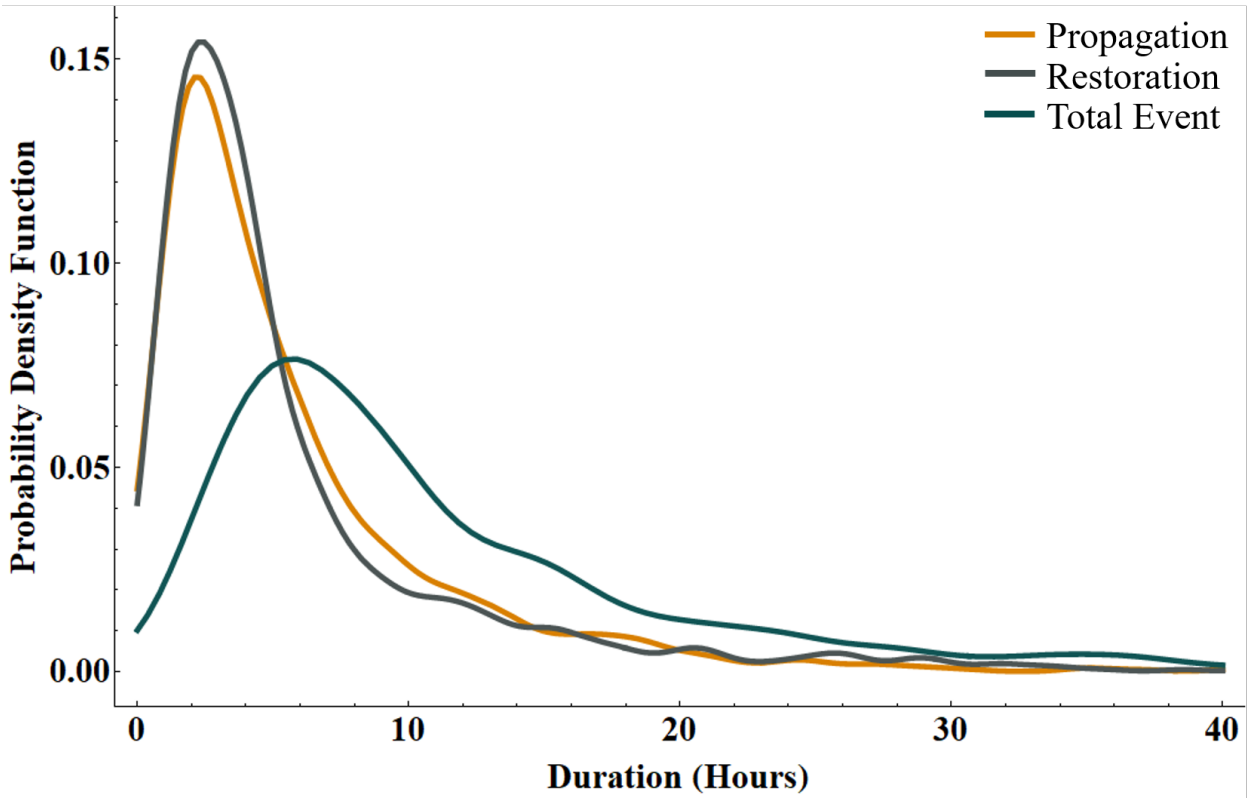

Figure 2.4: Distributions of propagation, restoration and total event durations.

2.2 are samples of these events. The large events had more outages caused by weather than the small and medium events. A common pattern of many large events is that outages caused by weather are followed by outages caused by tree limbs and other debris.

2.4.3 Nadir

The statistics for the resilience curve nadirs are summarized in Table 2.1. In Table 2.1, it can be seen that the utility can expect at least 5 simultaneously outaged components for any event. The average number of simultaneously outaged components expected in large events is over twice the average for medium events and over 8 times the average for small events.

Table 2.1: Nadir $C(t_N)$

Events	Mean	Median	Std.Dev.
Small	–5.25	–5	1.36
Medium	–13.81	–13	2.75
Large	–42.52	–34	22.13
All	–7.67	–5	9.59

2.4.4 Event Duration

The average event duration and their variabilities are summarized in Table 2.2, and the survival function of the event duration for each event size is shown in Fig. 2.5. The average event duration is 13 hours, but the variability is high. Small events are over within 24 hours with 95% confidence, but the corresponding upper bounds for the medium events are two times longer and the large events are four times longer. The utility will be able to make an assessment of the event duration for the event and also scale the estimate of the duration up or down if necessary for sudden changes to conditions.

Table 2.2: Event Duration (hours)

Events	Mean	Median	Std.Dev.	95%CI
Small	9.50	7.65	6.74	24
Medium	32.11	21.63	51.99	55
Large	49.48	41.42	23.31	99
All	13.07	8.5	19.07	36

2.4.5 Outage Propagation Process

The estimation and variability of the duration for the outage propagation process are shown in Table 2.3, and the survival function of the distribution of the propagation process for each group is in Fig. 2.6. The duration of propagation during any event is less than 18 hours with 95% confidence.

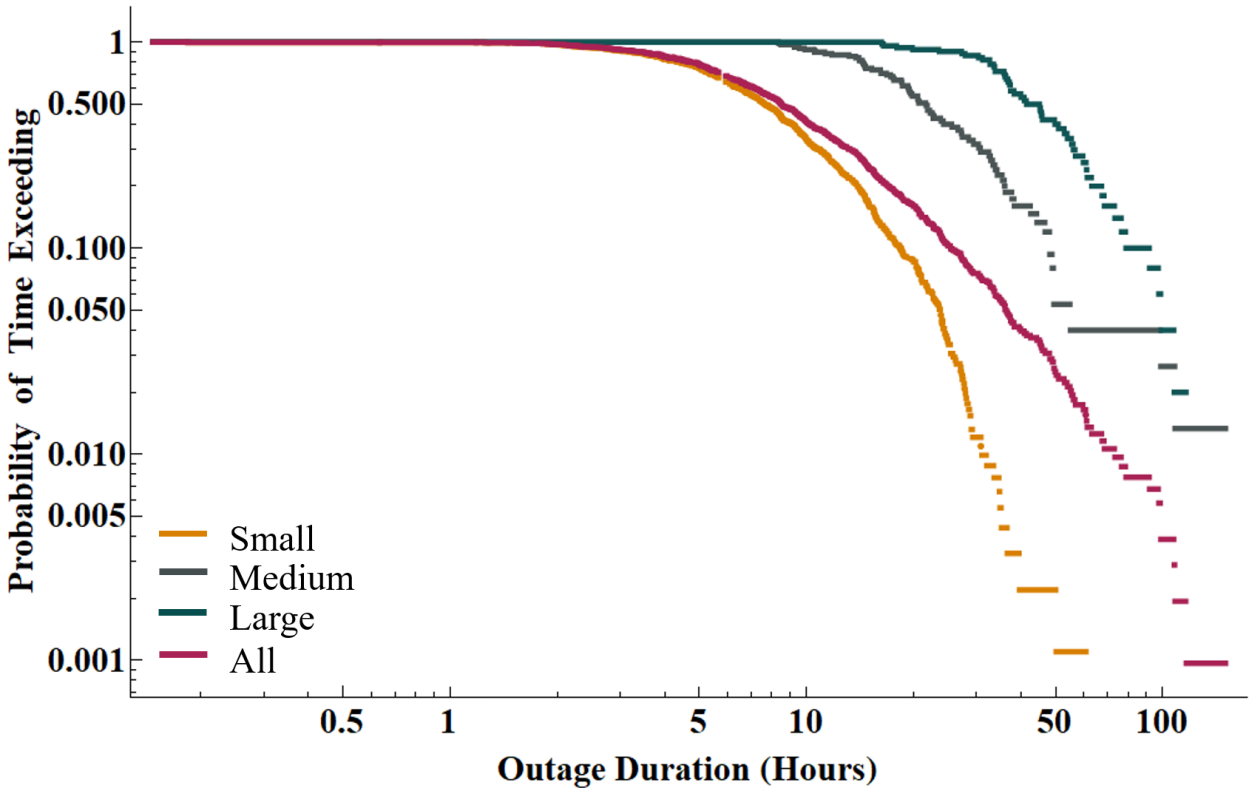

Figure 2.5: Survival functions of event duration.

2.4.6 Recovery Process

The estimation and variability of the duration for the recovery process are shown in Table 2.4 and the survival functions of the recovery process duration for each events size are shown in Fig. 2.7. The expected duration of recovery during any event is less than 22.6 hours with 95% confidence. This upper bound on duration for any event is about four times lower than the upper bound on duration for large events and twice that of small events.

Table 2.3: Propagation Process Duration (hours)

Events	Mean	Median	Std.Dev.	95%CI
Small	5.01	3.5	4.66	15
Medium	17.79	7.82	50.77	43
Large	14.78	10.92	11.91	40
All	6.41	3.93	15.06	18

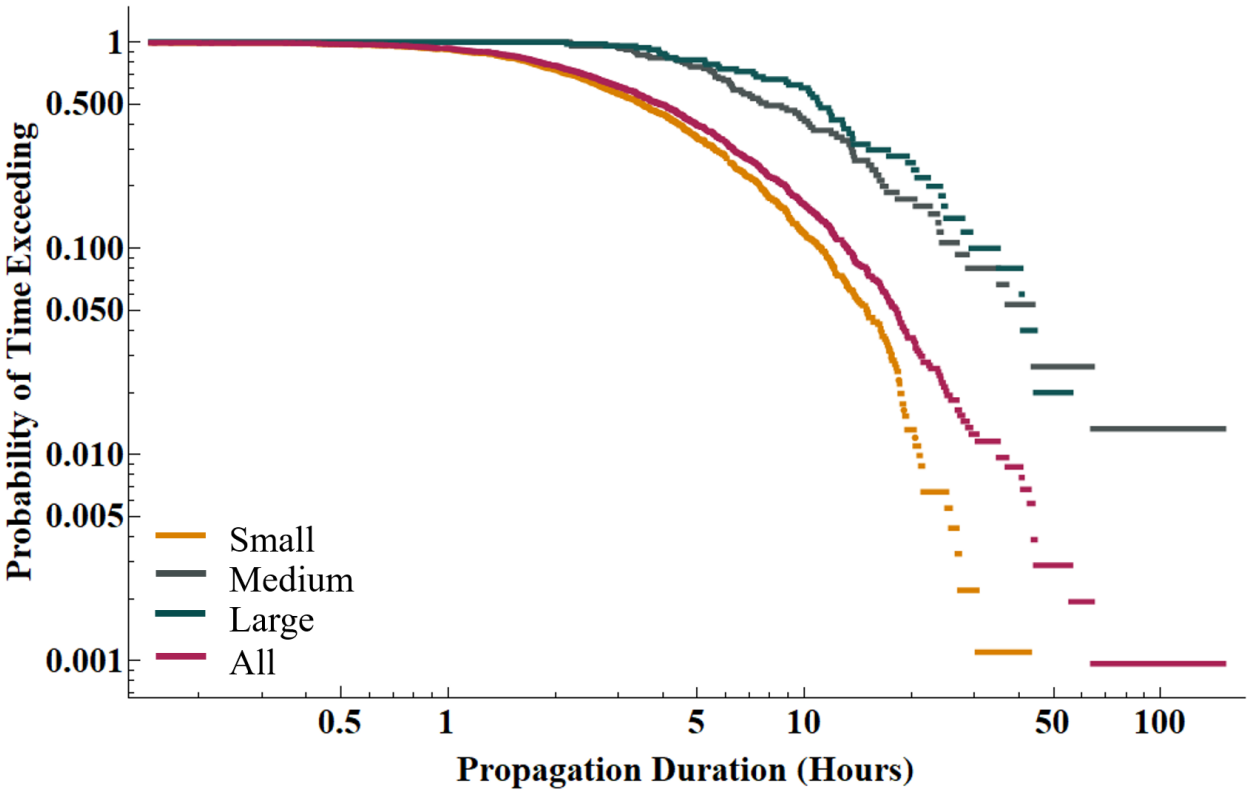

Figure 2.6: Survival functions of the propagation process duration.

Table 2.4: Restoration Duration (hours)

Events	Mean	Median	Std.Dev.	95%CI
Small	4.5	3.35	3.97	13
Medium	14.31	11.9	10.89	30
Large	34.7	28.37	21.77	85
All	6.67	3.73	9.58	22.6

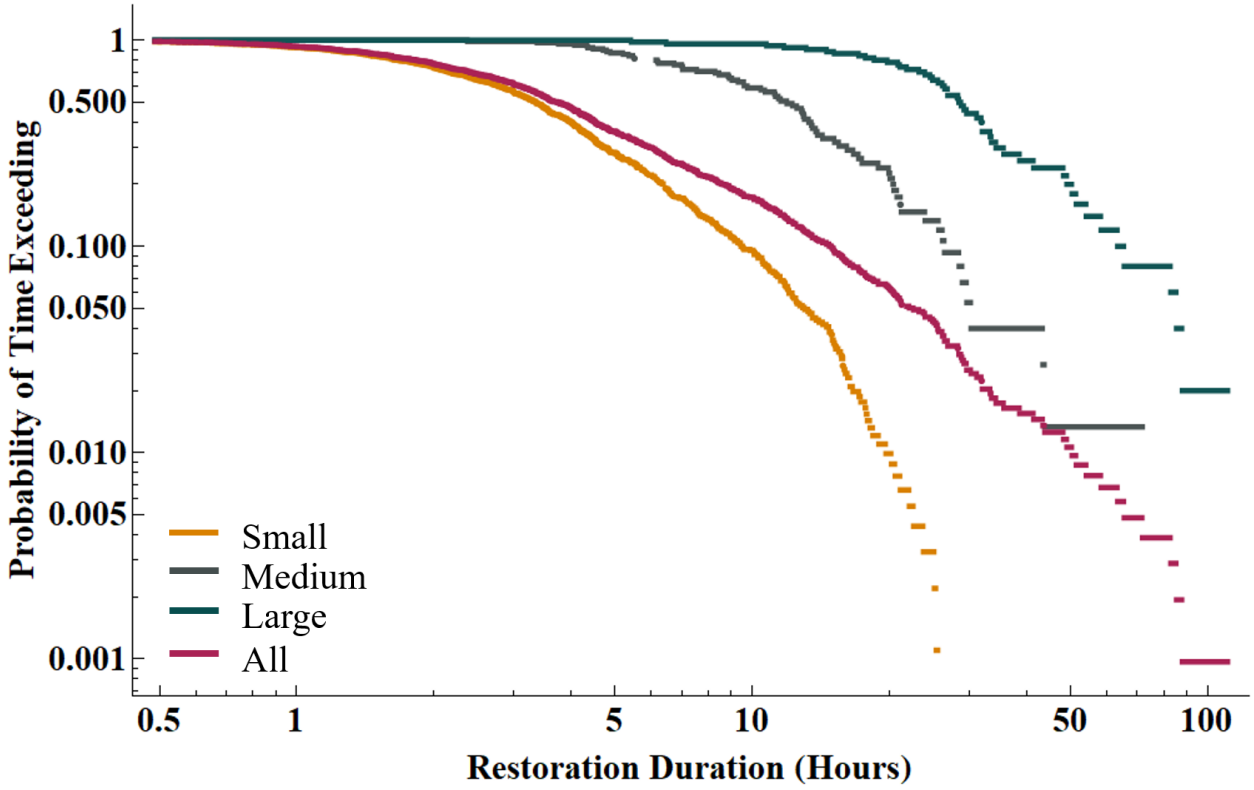

Figure 2.7: Survival functions of recovery process duration.

2.4.7 Customer Impact

The customer resilience curves for each event and the customer area under each curve were computed, and the statistics are shown in Table 2.5. Table 2.5 shows that for all events, the customer hours of outage are less than 2399 with 95% confidence.

Table 2.5: Customer Area (customer hours)

Events	Mean	Median	Std.Dev.	95%CI
Small	405	168	812	1466
Medium	1442	982	1944	5297
Large	2501	1972	204	5534
All	581	209	1145	2399

2.4.8 Inter-Arrival of Event Times

The rate at which the outages occur in time is known as inter-arrival times and the distribution of inter-arrival times for a Poisson process is an exponential. Our claim that the resilience curves detected are in fact NHPP is confirmed by the survival function of the distribution of the inter-arrival times of incidents within events. Figure 2.8 depicts the distribution of time between occurrences inside the resilience curves as exponential, revealing that there are two unique rates of *lambda* for all categories and all detected curves combined. The higher rate for smaller time differences is evidence of bunching together in time for some of the outages. The higher rates is particularly evident for the larger outages.

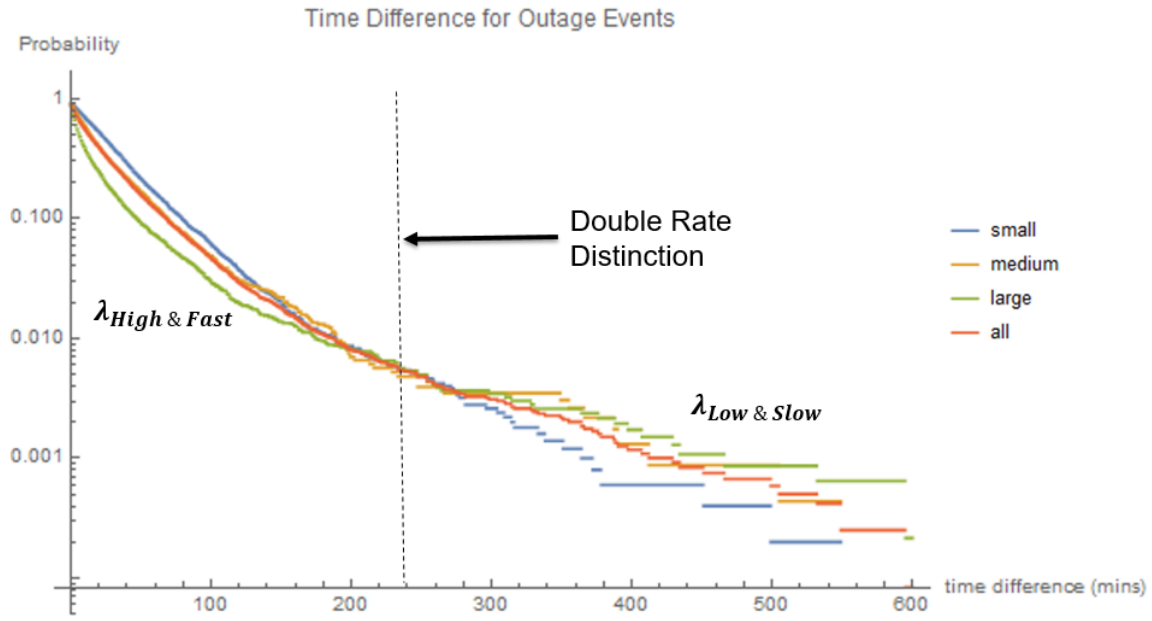

Figure 2.8: The λ_1 relates to the outage events that have higher probability and fast in-between times and λ_2 represents events with lower probability and larger times in-between.

The two distinct are separated by the dash line in Figure 2.8 indicates that there is one process with a fast rate of outages ($\lambda_{High\&Fast}$) occurring at a high probability with very short times in between outages and a process with a very slow rate ($\lambda_{Low\&Slow}$) long time in between outages

with a low probability. Approximating $\lambda_{Low\&Slow}$ is useful to the understanding the resiliency of distribution systems to extreme weather and identifying vulnerabilities to the infrastructure.

2.4.9 Average Outage and Recovery Rates

The average recovery process rate and the average outage process rate are calculated from the resilience triangles by dividing the magnitude of the nadir by the duration of the process. Table 2.6 shows that for medium and large events, the average recovery rate is slower than the average outage rate. The 95% confidence interval for the average recovery rate shown in Table 2.6 is a one-sided lower confidence interval. That is, the probability that the average recovery rate is more than the given value is 0.95. After the nadir of a resilience event, when the damage has been inspected and the current number of outages is known, the average recovery rate and its 95% confident lower bound can be multiplied by the number of outages to estimate the expected recovery time and its 95% confident upper bound recovery time.

Table 2.6: Average outage and recovery rates (per hour)

	Average outage rate			Average recovery rate			
Events	Mean	Median	StdDev	Mean	Median	StdDev	95%CI
Small	0.48	0.68	0.42	0.45	0.66	0.35	0.37
Medium	0.50	0.55	0.70	0.68	0.97	0.95	0.22
Large	0.19	0.32	0.18	0.63	0.66	0.95	0.27
All	0.45	0.66	0.38	0.47	0.67	0.37	0.37

2.5 Conclusions

This chapter systematically detects and extracts resilience curves from 5 years of distribution utility outage data for each resilience event in which outages accumulate. For each event, the resilience curve is divided into an outage propagation period and a recovery period using the

nadir of the resilience curve. The events are classified into small, medium, and large, and the statistics of resilience metrics such as the durations of the event, propagation and recovery and the customer hours lost are computed for each size of event. In contrast to previous work, we compute statistics of groups of typical resilience events rather than focusing on single resilience events. The statistics for the average durations and their variability should be helpful in estimating event and recovery times for future events before or while they are occurring. This first analysis spurred the development of better ways to extract the metrics in Chapter3, and further exploration of the causes and weather in Chapter 4.

2.6 References

[1] N. K. Carrington, S. Ma, I. Dobson, and Z. Wang, "Extracting resilience statistics from utility data in distribution grids," in *2020 IEEE Power Energy Society General Meeting (PESGM)*, 2020, pp. 1–5.

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

[3] M. Ouyang, L. Dueñas-Osorio, and X. Min, "A three-stage resilience analysis framework for urban infrastructure systems," *Structural safety*, vol. 36, pp. 23–31, 2012.

[4] Y. Wei, C. Ji, F. Galvan, S. Couvillon, G. Orellana, and J. Momoh, "Non-stationary random process for large-scale failure and recovery of power distributions," *Applied Mathematics*, 2012.

[5] M.-Y. Chow, L. S. Taylor, and M.-S. Chow, "Time of outage restoration analysis in distribution systems," *IEEE Transactions on Power Delivery*, vol. 11, no. 3, pp. 1652–1658, 1996.

[6] B. M. Ayyub, "Systems resilience for multihazard environments: Definition, metrics, and valuation for decision making," *Risk Analysis*, vol. 34, no. 2, pp. 340–355, 2014.

CHAPTER 3. EXTRACTING RESILIENCE METRICS FROM DISTRIBUTION UTILITY DATA USING OUTAGE AND RESTORE PROCESS STATISTICS

Nichelle'Le K. Carrington, Ian Dobson, and Zhaoyu Wang, Department of Electrical and Computer Engineering Iowa State University, Ames, Iowa, USA

Modified from a manuscript published in November 2021 issue of *IEEE Transactions on Power Systems* [1]

3.1 Abstract

Resilience curves track the accumulation and restoration of outages during an event on an electric distribution grid. We show that a resilience curve generated from utility data can always be decomposed into an outage process and a restore process and that these processes generally overlap in time. We use many events in real utility data to characterize the statistics of these processes, and derive formulas based on these statistics for resilience metrics such as restore duration, customer hours not served, and outage and restore rates. Estimating the variability of restore duration allows us to predict a maximum restore duration with 95% confidence.

3.2 Overview

The frequency of outages fluctuates given the environmental conditions that they occur under. Under normal conditions, component outages in electric power distribution systems occur at a low rate and are restored as they occur. However, when there is severe weather or other extreme stresses, the component outages occur at a high rate, and the outaged components accumulate until they are gradually restored. Chapter 3 idealized restores and outage processes without overlap. In contrast, this chapter addresses and solves the overlapping of outages and restore process.

We use the resilience curves found in the distribution outage data using the distribution system data from section 1.3.1 for the work in this chapter.

In practice it is customary to divide resilience curves into successive, non-overlapping phases. For example, Nan [2] describes a disruptive outaging phase followed by a recovery phase,while Panteli [3] and Ouyang [4] describes a resilience trapezoid with the three phases of progressive disturbances, then a degraded or assessment phase, then recovery in a resilience triangle. Similarly, Yodo [5] describes resilience curves in terms of successive unreliability, disrupted, and recovery phases with triangles, trapezoids, and other curves. Carrington [6] uses the nadir of resilience curves from utility data to separate an outage phase from a recovery phase. Many other papers have similar accounts of resilience phases as a model to extract resilience metrics. These distinct phases of resilience are conceptually compelling. Moreover, the dimensions, slopes, and areas of the resilience triangles and trapezoids define standard resilience metrics of event duration, average rates of outage or recovery, and overall impact.

However, working with our utility data suggests a different point of view in which outage and restore processes routinely overlap in time[1]. That is, for practical processing of real distribution system outage data our method decomposes the resilience curve not into successive phases but into an outage and a restore processes that occur together for duration of the event.

In this chapter the ability to obtain resilience metrics from decomposed processes will be demonstrated. The aim of this approach is to convey that outage and restore processes can always be combined into a resilience curve, and any resilience curve can always be decomposed into an outage process and a restore process. The outage process is statistically characterized by the times between successive outages or the outage rate.The restore process starts after a delay and is statistically characterized by the times between successive restores or the restore rate. The duration of the restore process and of the entire event are then easily obtained standard metrics.

[1]The average fraction of event duration for which outage and restore processes overlap (average of $(o_n - r_1)/(r_n - o_1)$ in the notation of section 3.3) is 0.61 for events with 10 to 20 outages, 0.89 for events with 100 to 200 outages, and 0.95 for events with 1000 to 2000 outages.

There is also a conventional customer resilience curve tracking the number of customers out during the event that we also obtain from the utility data. This customer resilience curve can also be decomposed into a customer outage process and a customer restore process. We can measure the impact of an event by the customer hours lost, which is the area under the customer resilience curve and a well-known resilience metric [2, 3, 5]. We compute the mean customer hours lost from the statistics of the customer outage and restore processes.

Previous pioneering work on queueing models of reliability and resilience has used outage and restore processes. Zapata [7] models distribution system reliability with outages as a point process arriving at a queue that is serviced by a repair process with multiple crews to produce an output that is a restore process. Wei and Ji [8] analyze distribution system resilience to particular severe hurricanes with an outage process arriving at a queue with a repair process to produce a restore process. In [8], these processes vary in both time and space as the hurricane progresses.Both [7] and [8] statistically model the outage process and the repair process of components, and then calculate the restore process. With our focus on studying the overall system resilience with real data, we can model the restore process directly from the data, and avoid the complexities of explicitly modeling the repair of components and assuming an order in which they are repaired.

Our methods require a sufficient number of events for good statistics, so that this chapter addresses the more common, less extreme events. Therefore there is little overlap of this paper with [8], which addresses individual instances of the most extreme events (direct hits by a hurricane), for which there are few events for a given utility.

There are several methods of estimating the number of outages in an anticipated storm [9–15], including practical utility application in [15]. Since some of our statistics show a dependence on the number of outages, this capability to predict the number of outages will be useful in applying the results in this chapter to anticipated storms.

The utility data also yields estimates of the variability of the outage and restore processes, enabling estimates of the variability of the restore duration. Then, given the estimate of the num-

ber of outages, we can use our restore time statistics to predict upper bounds of the restore duration of an anticipated storm, such as its 95th percentile. The upper bound is intended to help the utility predict when the restore process will be completed with more confidence.

With a similar overall aim, previous work estimates individual component restoration times from utility data in different ways. For example, Jaech [16] predicts a gamma distribution of individual component outage restoration times and customer hours lost with a neural network that processes utility records and wind speed, outage time and date, and hence obtains upper bounds of individual component restoration times. Chow [17] analyzes the contributions of timing, faults, protection, outage types and weather to individual component restoration times. Liu [18] fit generalized additive accelerated failure time models to hurricane and ice storm utility data. The individual outages were then combined to give system restoration curves at the county level.

With these innovations, we are able to compute standard metrics for resilience events from practical utility data and evaluate customer impact risks as a function of size.

3.3 Outage and restore processes

The resilience events of interest occur when outaged components accumulate before being restored. Each event has a conventional resilience curve $C(t)$ for component outages. $C(t)$ is the negative of the cumulative number of component outages as a function of time t. For example, the orange curve at the bottom portion of Figure 3.1 shows $C(t)$ for an event with 10 component outages. We now explain the outage and restore processes and how they relate to the resilience curve.

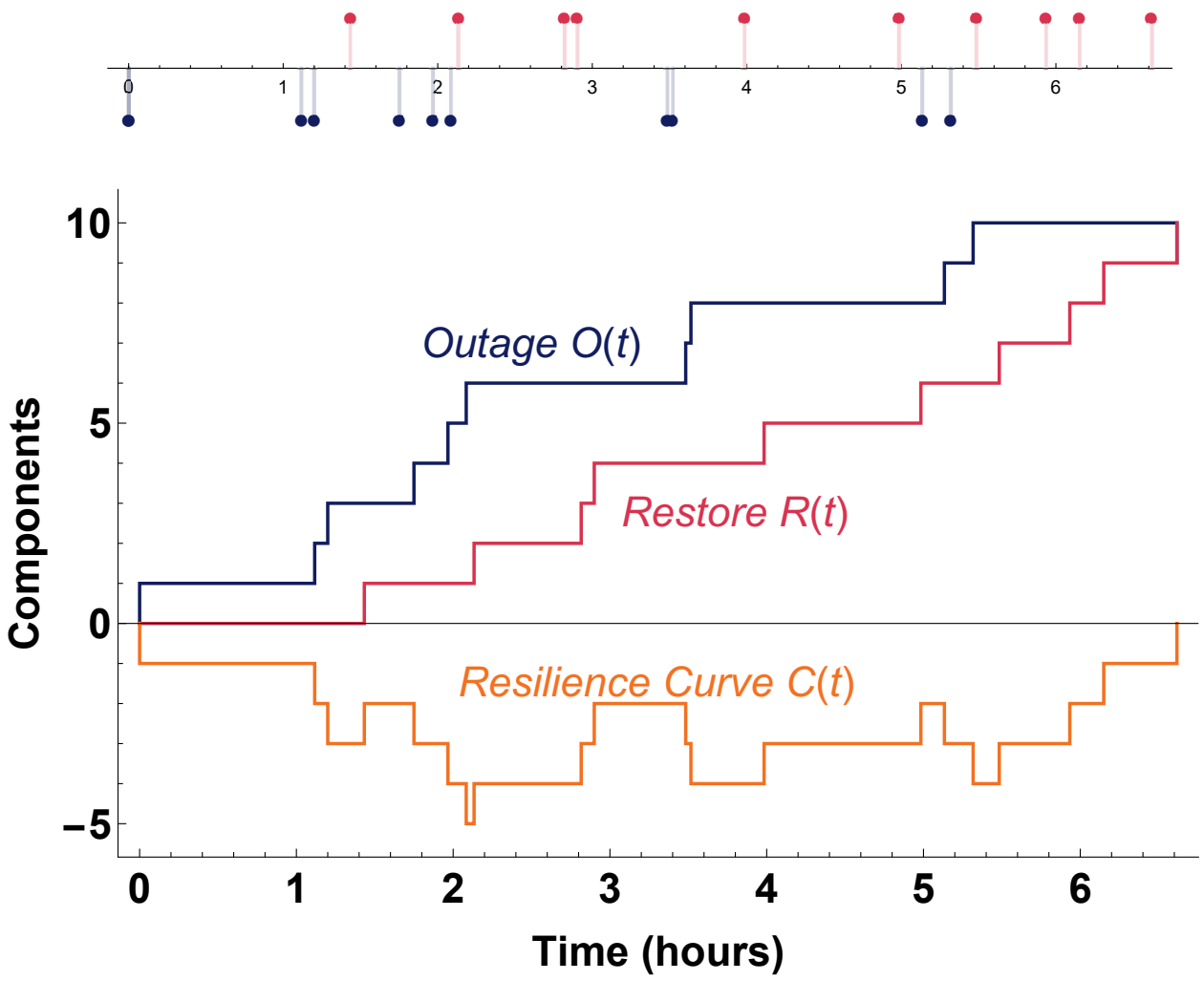

Figure 3.1: A component resilience curve and its associated outage and restore processes.

3.3.1 Examples of component outage and restore processes

We start with no components outaged. Then 10 components outage at times $o_1 \leq o_2 \leq ... \leq o_{10}$ as shown by the tick marks below the top time line of Figure 3.1. The restore times are $r_1 \leq r_2 \leq ... \leq r_{10}$ as shown by the tick marks above the top time line of Figure 3.1. The restore times are numbered in the time order that they occur. The cumulative number of outages $O(t)$ at time t and the cumulative number of restores $R(t)$ at time t are defined by counting 1 for each outage

or restore before time t:

$$O(t) = \sum_{k \text{ with } o_k \leq t} 1 \tag{3.1}$$

$$R(t) = \sum_{k \text{ with } r_k \leq t} 1 \tag{3.2}$$

Figure 3.1 shows the cumulative number of outages $O(t)$ and the cumulative number of restores $R(t)$. In this case, $O(t)$ and $R(t)$ increase from zero to the total number of outages 10.

The cumulative number of component outages at time t is $O(t) - R(t)$. The component resilience curve $C(t)$ is defined as the negative of the cumulative number of component outages at time t so that

$$C(t) = R(t) - O(t) \tag{3.3}$$

Figure 3.1 shows how the resilience curve $C(t)$ can be decomposed into the restore process minus the outage process.

It is clear from (3.3) that any outage and restore processes $O(t)$ and $R(t)$ define a resilience curve $C(t)$. Moreover, any resilience curve $C(t)$ can be uniquely decomposed as (3.3) into outage and restoration processes $O(t)$ and $R(t)$ that increase from zero to the number of outaged components n. In mathematics this decomposition is known as the Jordan decomposition [19] of functions of bounded variation[2].

There is noticeable variety in the forms of the component resilience curves in our utility data. The examples (except for the first example) in Figure 3.2, show these curves and their decompositions into outage and restore processes.

[2]The total variation of $C(t)$ is $2n$, which is bounded. In our case the Jordan decomposition (3.3) is minimal and unique since we require that $O(t) = R(t) = 0$ for $t < o_1$ and $O(t) + R(t) = 2n$ for $t > r_n$ [20, defn. 2.4(2), thm. 2.5(2)], [21, sec. 9-4]. Since $C(t)$ is minus the cumulative number of outages, if there are simultaneous restores and outages, for example, m_r restores and m_o outages all occurring precisely at time t, then only their difference $m_r - m_o$ contributes to $C(t)$, and we assume that only $|m_r - m_o|$ contributes to the total number n of outages or restores. Note that $O(t), R(t), C(t)$ are right continuous.

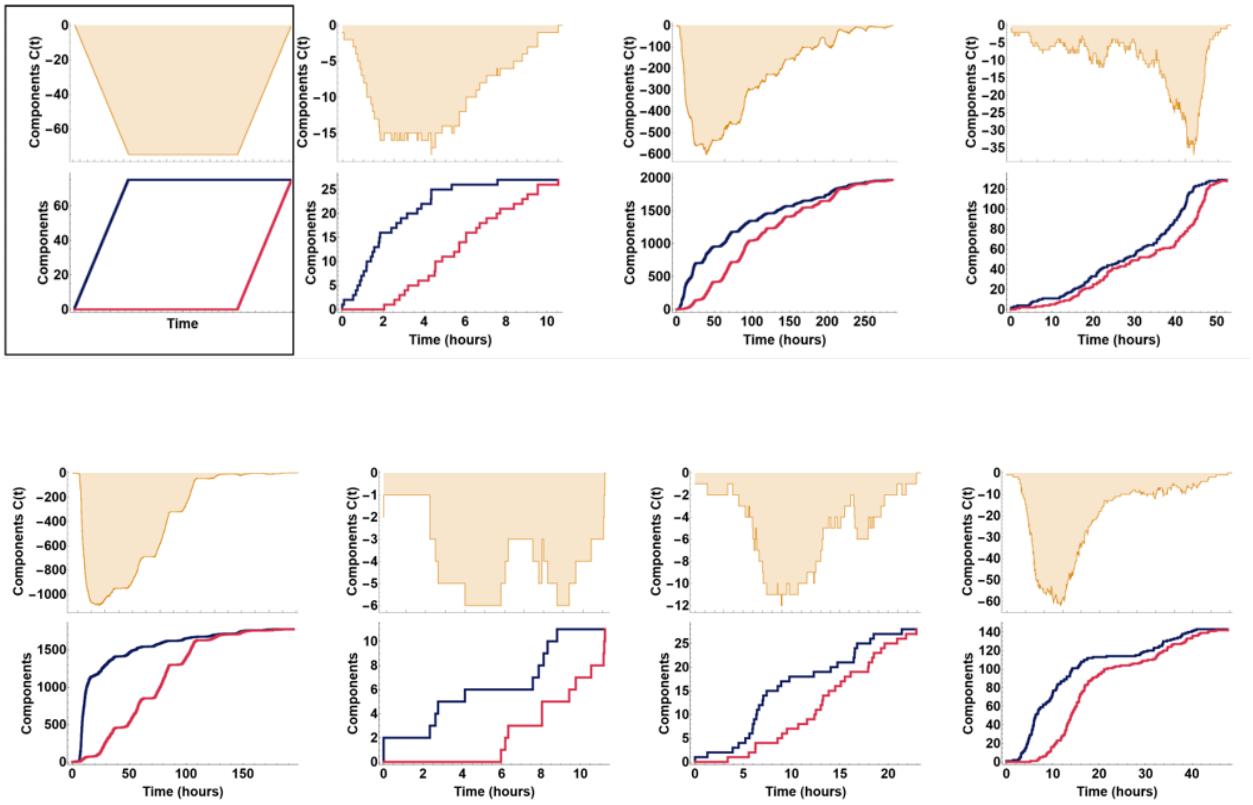

Figure 3.2: Component resilience curves (upper rows with one shaded curve) and their corresponding decompositions into outage and restore processes (lower rows with blue and red curves). The first example is an idealized case with trapezoidal resilience curve and all the rest are examples from utility data.

For comparison, the first example in Figure 3.2 shows a conventional, idealized case of a trapezoidal resilience curve.

Considering the outage and restore processes separately is useful because they correspond to different aspects of system resilience: the outage process results from individual component strengths under bad weather stress and the restore process results from the number and performance of restoration crews and automated or remote switching.

3.3.2 Extracting events from utility data

The historical distribution outage data described in 1.3.1 was used in the work for this chapter. The method detailed in 2.3.1 was applied to define events. By definition,the start of an event is defined by an initial outage that occurs when all components are functional, and the end of the same event is defined by the first subsequent time when all the components are restored. That is, the event starts when the cumulative number of failures $C(t)$ first changes from zero and ends when $C(t)$ returns to zero. Applying this event processing to the historical data yields 2618 events.

The component and customer resilience curves for each event were decomposed into outage and restore processes as explained in more detail in the next subsection.

In particular, we sort the combined component outage and restore times by their order of occurrence and then calculate the cumulative number of outages $C(t)$ at all the outage and restore times. Each restore at time r_n for which $C(r_n) = 0$ is the end of an event and the immediately following outage is the start of the next event. Note that if the event has n outages, then it must have n restores to allow the cumulative number of outage $C(t)$ to return to zero at time $t = r_n$.

Applying this event processing to the historical data yields 2618 events. The component and customer resilience curves for each event are decomposed into outage and restore processes as explained in more detail in the next subsection.

3.3.3 Component outage and restore processes

This subsection explains the outage and restore processes in more detail and shows how their statistics are extracted from the events in the utility data.

Suppose that $o_1 \leq o_2 \leq ... \leq o_n$ are the component outage times in an event in order of occurrence and that $\Delta o_k = o_{k+1} - o_k$, $k = 1, ..., n-1$ are the times between successive component outages.

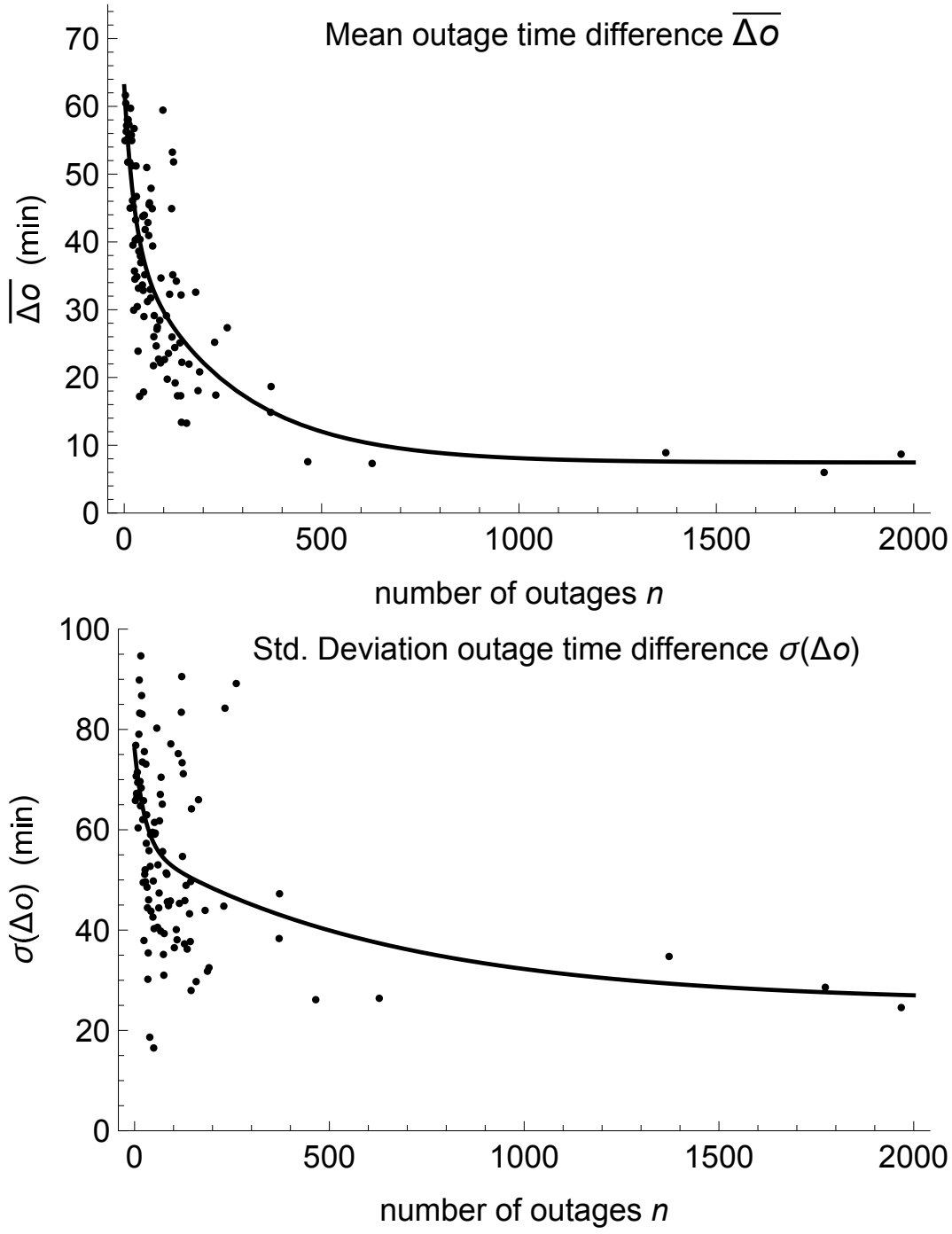

Figure 3.3: Mean and standard deviation of outage time difference empirical data (dots) and fitted curve as a function of number of outages n.

The outage time differences Δo_k, $k = 1, ..., n-1$ can be regarded as independent samples from a probability distribution of outage time differences Δo. We want to find the mean and standard deviation of Δo as a function of n.

To do this, we combine the outage time differences Δo_k for all the events with n outages and calculate their mean and standard deviation. Then we fit the empirical mean and the standard deviation as functions of n with a linear combination of a constant and 2 exponential functions to smooth and interpolate the data as shown in Figure 3.3. The functional fits are

$$\overline{\Delta o} = 7.45 + 23.3e^{-0.0388n} + 32.2e^{-0.00391n} \quad \text{min} \tag{3.4}$$

$$\sigma(\Delta o) = 25.6 + 19.5e^{-0.0375n} + 30.9e^{-0.00153n} \quad \text{min} \tag{3.5}$$

Suppose that $r_1 \leq r_2 \leq ... \leq r_n$ are the component restore times in order of occurrence. Note that the component outaged in the kth outage can be different from the component restored in the kth restore. In effect, we disregard *which* component is restored and only track that *some* component is restored. (We call $r_1, r_2, ..., r_n$ component restore times to minimize any confusion with the restoration or repair times of particular components.)

Let $\Delta r_k = r_{k+1} - r_k$, $k = 1, ..., n-1$ be the times between successive component restores. The restore time differences Δr_k, $k = 1, ..., n-1$ can be regarded as independent samples from a probability distribution of restore time differences Δr. We extract the statistics of Δr from the utility data as a function of the number of outages n similarly as Δo. Figure 3.4 plots the mean restore time difference $\overline{\Delta r}$ and the standard deviation of the restore time difference $\sigma(\Delta r)$ and the functions fitted. The functional fits are

$$\overline{\Delta r} = 7.64 + 30.8e^{-0.0514n} + 33.8e^{-0.00391n} \quad \text{min} \tag{3.6}$$

$$\sigma(\Delta r) = 35.3 + 43.7e^{-0.0224n} \quad \text{min} \tag{3.7}$$

Let $\Delta r_0 = r_1 - o_1$ be the delay in the start of the restoration process relative to the start of the event at time o_1. One factor contributing to Δr_0 is utility inspection crews and clean-up crews working to ensure the safety of the area and assess the damage needing repair. There is no clear

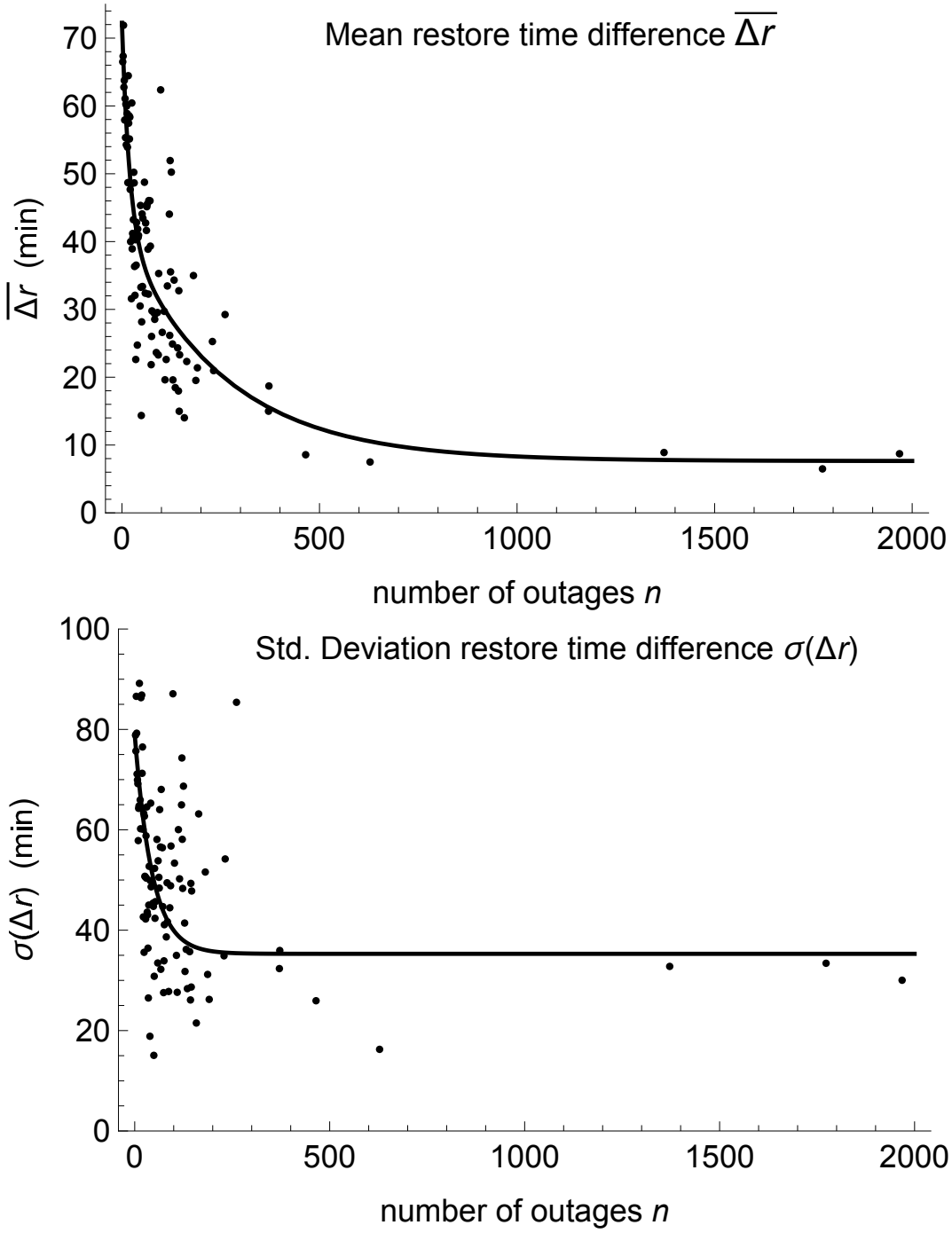

Figure 3.4: Mean and standard deviation of restore time difference empirical data (dots) and fitted curve as a function of number of outages n.

trend in variation of Δr_0 with n, so we combine the data for Δr_0 for all events with 2 or more outages and compute

$$\overline{\Delta r_0} = 132 \text{ min} \qquad \text{and} \qquad \sigma(\Delta r_0) = 92.4 \text{ min} \tag{3.8}$$

3.3.4 Customer outage and restore processes

The utility data records the number of customers outaged by each component outage, allowing us to similarly analyze resilience curves tracking the number of customers out and outage and restore processes for the number of customers. Generalizing (3.1) and (3.2), the cumulative numbers of customers out $O^{\text{cust}}(t)$ and customers restored $R^{\text{cust}}(t)$ at time t are

$$O^{\text{cust}}(t) = \sum_{k \text{ with } o_k \leq t} c_k^{\text{out}} \tag{3.9}$$

$$R^{\text{cust}}(t) = \sum_{k \text{ with } r_k \leq t} c_k^{\text{res}} \tag{3.10}$$

The customer resilience curve $C^{\text{cust}}(t)$ is now obtained similarly to the component resilience curve (3.3) as

$$C^{\text{cust}}(t) = R^{\text{cust}}(t) - O^{\text{cust}}(t). \tag{3.11}$$

Figure 3.5 shows an example of customer processes $O^{\text{cust}}(t)$ and $R^{\text{cust}}(t)$ and the resilience curve $C^{\text{cust}}(t)$. The processes $O^{\text{cust}}(t)$ and $R^{\text{cust}}(t)$ increase from zero to the total number of customers out, which is 181 in this example.

When an outage that disconnected customers is restored, the same number of customers are restored. However, outages are not necessarily restored in the order that the outages occurred. Therefore the numbers of customers restored $c_1^{\text{res}}, c_2^{\text{res}}, ..., c_n^{\text{res}}$ are a permutation of the numbers of customers out $c_1^{\text{out}}, c_2^{\text{out}}, ..., c_n^{\text{out}}$.

The numbers of customers out $c_1^{\text{out}}, c_2^{\text{out}}, ..., c_n^{\text{out}}$ can be regarded as n independent samples from a distribution c of number of customers out. The customers restored $c_1^{\text{res}}, c_2^{\text{res}}, ..., c_n^{\text{res}}$ can also be regarded as n independent samples from c. We combine the data for customers out for all events[3]

[3] 2% of the customer data are blank entries that we replaced with 0.

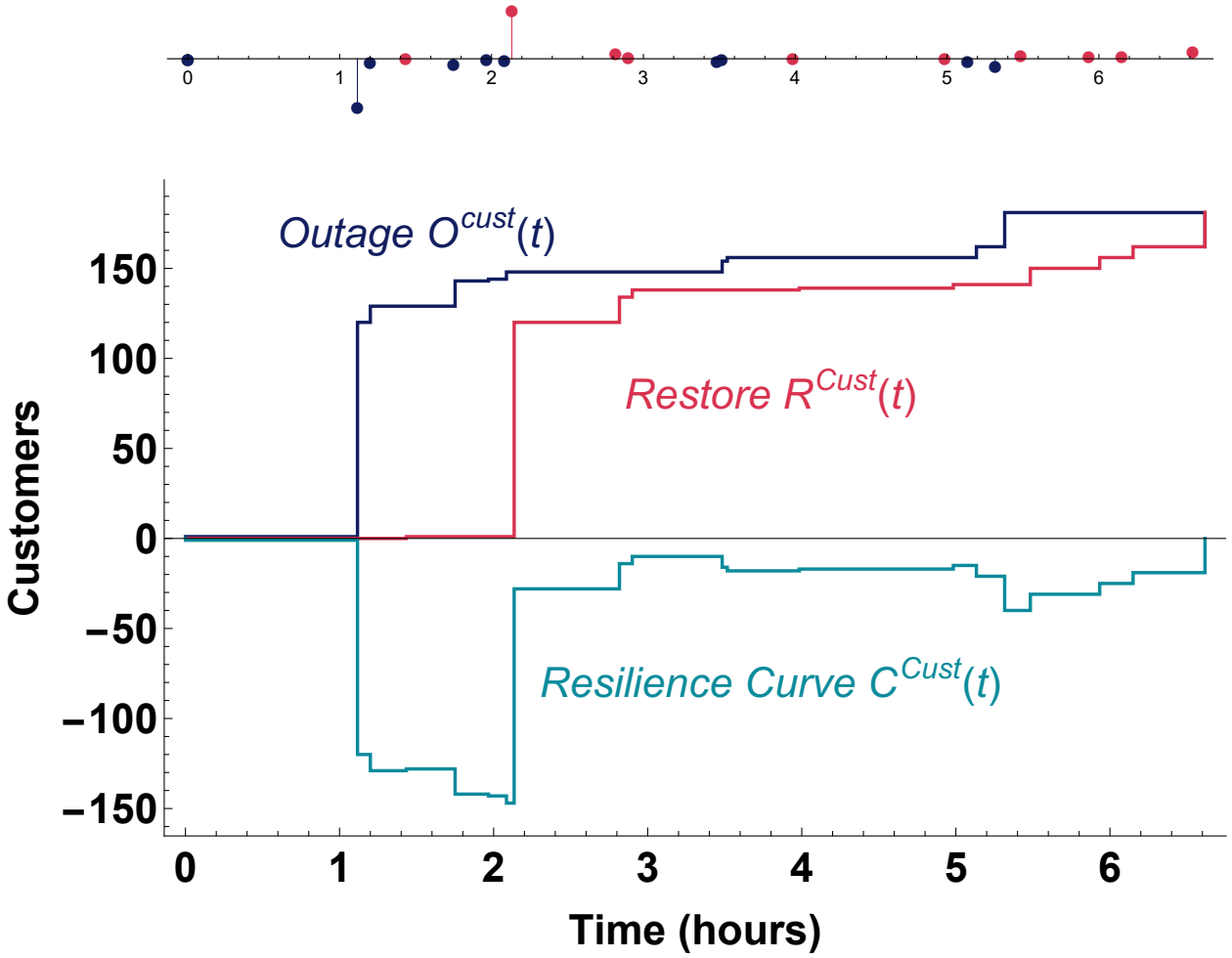

Figure 3.5: A customer resilience curve and its associated customer outage and customer restore processes for the same event as Figure 3.1.

and compute the mean and standard deviation of the number of customers:

$$\overline{c} = 54.0 \qquad \text{and} \qquad \sigma(c) = 180. \tag{3.12}$$

The customers out for each outage are determined by the location of the outage in the network, and the distribution of the number of customers out is determined by the overall network design and its vulnerabilities.

3.4 Resilience metrics

We express event durations in terms of the time differences of the restore process and then derive formulas for the mean and standard deviations of the restore duration and the event duration. We also combine the time differences with the customers outage statistics to derive a formula for the mean customer hours lost. The average outage and restore rates are obtained.

The restore process starts at time r_1 and ends at time r_n, so the restore duration is

$$\begin{aligned} D_R &= r_n - r_1 = (r_2 - r_1) + (r_3 - r_2) + ... + (r_n - r_{n-1}) \\ &= \Delta r_1 + \Delta r_2 + ... + \Delta r_{n-1} \end{aligned} \tag{3.13}$$

The mean restore duration is then

$$\overline{D_R} = (n-1)\overline{\Delta r} \tag{3.14}$$

Assuming that Δr_1, Δr_2, ..., Δr_{n-1} are independent, we obtain $\sigma^2(D_R) = (n-1)\,\sigma^2(\Delta r)$ and

$$\sigma(D_R) = \sqrt{n-1}\,\sigma(\Delta r) \tag{3.15}$$

In (3.13), the restore duration D_R is measured until the last restore of the event. If it is preferred to measure the restore duration until, say, 95% of the outages are restored, then this can easily be done by replacing $n-1$ in (3.14) and (3.15) by $\lceil 0.95\,n - 1 \rceil$, where the ceiling function $\lceil \cdot \rceil$ rounds up to the nearest integer.

Each event starts at the first outage time o_1 and ends at the last restore time r_n. Then the event duration is

$$D_E = r_n - o_1 = (r_1 - o_1) + (r_n - r_1) = \Delta r_0 + D_R \tag{3.16}$$

Using (3.14), the mean event duration is

$$\overline{D_E} = \overline{\Delta r_0} + (n-1)\overline{\Delta r} \tag{3.17}$$

and, since Δr_0 and D_R are independent, we use (3.15) to obtain $\sigma^2(D_E) = \sigma^2(\Delta r_0) + (n-1)\sigma^2(\Delta r)$ and

$$\sigma(D_E) = \sqrt{\sigma^2(\Delta r_0) + (n-1)\sigma^2(\Delta r)} \tag{3.18}$$

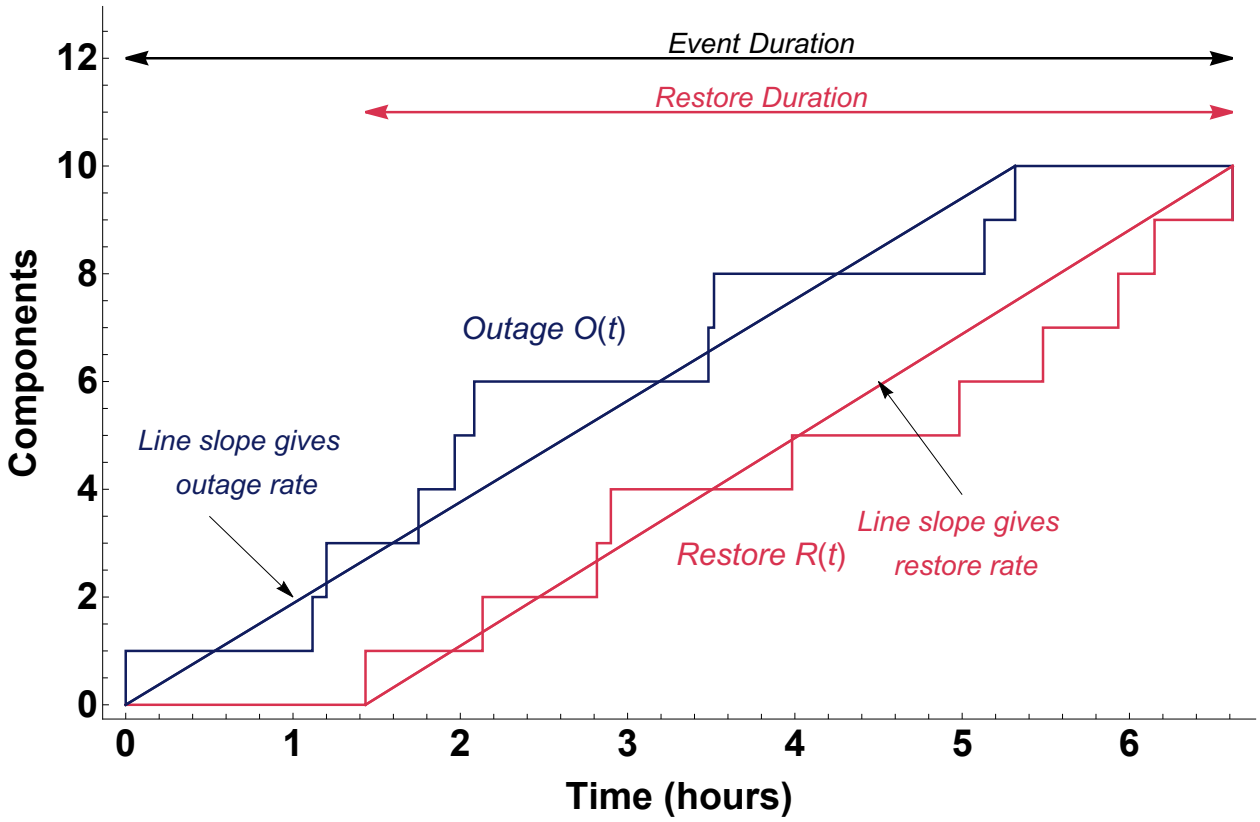

Figure 3.6: Resilience metrics (durations and rates) for the component outage and restore processes.

The restore and outage rates during events[4] are

$$\lambda_R = (\overline{\Delta r})^{-1} \tag{3.19}$$

$$\lambda_O = (\overline{\Delta o})^{-1} \tag{3.20}$$

[4]The outage rate measured over a year is much lower than the outage rate during events because it accounts for the time between events.

In (3.19) we obtain the restore rate λ_R from $\overline{\Delta r}$, which is a quantity averaged over events. This restore rate λ_R should be distinguished from the instantaneous restore rate $\lambda_R^{\text{inst}}(t)$, which has been observed in [8] to vary with time for the largest events.

The restore rate λ_R averaged over events can usefully apply even as the instantaneous restore rate varies.[5]

The restore process depends largely on the restoration capability available to the utility. The restore process metrics are $\overline{\Delta r_0}$ and $\overline{D_R}$ or λ_R. Increasing the number of utility crews or their effectiveness would decrease $\overline{\Delta r_0}$ and $\overline{D_R}$, and increase λ_R.

The outage process depends on a combination of the weather impact and the condition and strength of the grid components. The number of outages n will vary with the weather and the condition of the grid components, increasing if the weather is more extreme or more prolonged,

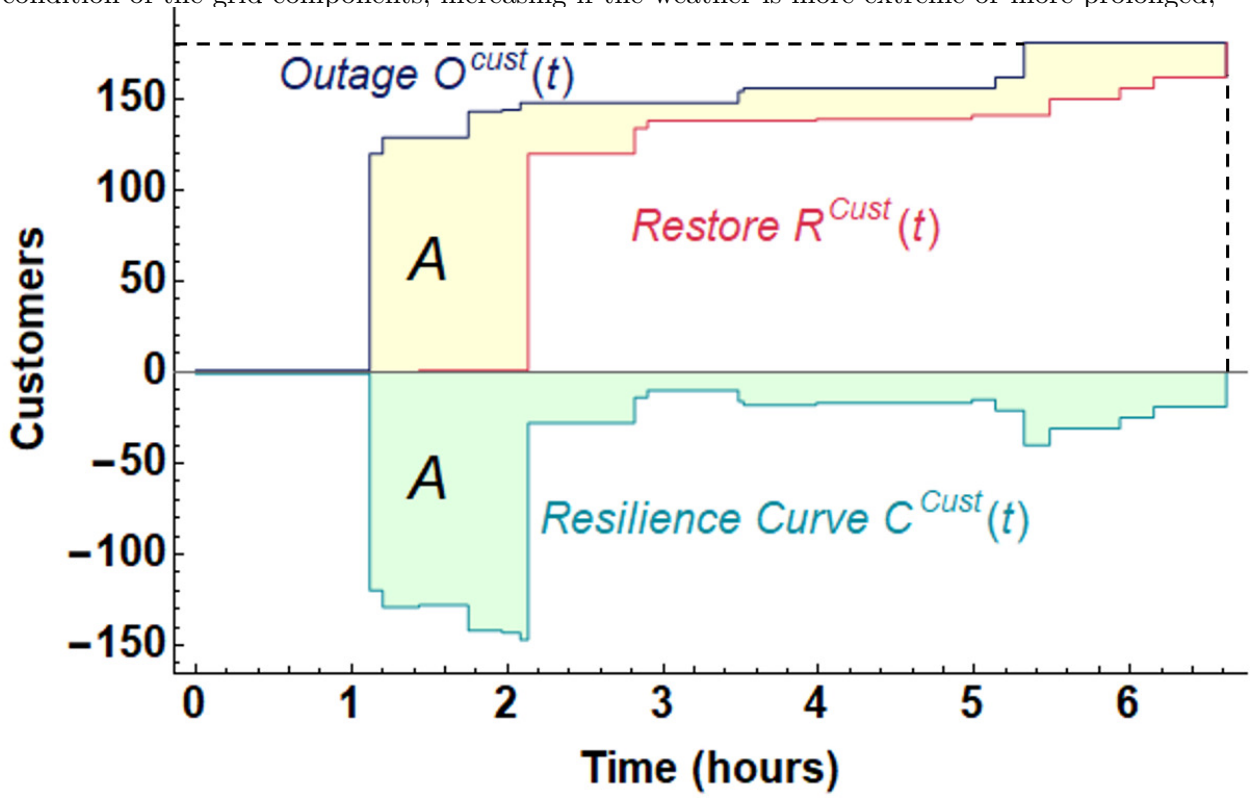

Figure 3.7: Area A under resilience curve is the customer hours metric and is equal to the area A between the outage and restore processes.

[5]Consider the idealized case of a non homogeneous Poisson recovery process. Suppose there are 3 restoring processes on the time intervals $D^{(1)}$, $D^{(2)}$, $D^{(3)}$, each of duration T and with n restores. Then the expected value of $\int_{t\in D^{(i)}} \lambda_R^{\text{inst}}(t)dt$ is $n-1$ and the duration $T = \sum_{k=1}^{n-1} \Delta r_k^{(i)} = (n-1)\overline{\Delta r}^{(i)}$ for $i = 1,2,3$. Then we estimate $\lambda_R = \frac{1}{3T}\sum_{i=1}^{3}\int_{t\in D^{(i)}} \lambda_R^{\text{inst}}(t)dt$ as $\frac{3(n-1)}{(n-1)(\overline{\Delta r}^{(1)}+\overline{\Delta r}^{(2)}+\overline{\Delta r}^{(3)})} = (\overline{\Delta r})^{-1}$.

We use the customer hours lost A to quantify the customer impact of an event. A is the area under the customer resilience curve:

$$A = -\int_{o_1}^{r_n} C^{\text{cust}}(t)dt \tag{3.21}$$

The minus sign in (3.21) makes A a positive area. Using (3.11), A is also the area between the customer outage and restore curves:

$$A = \int_{o_1}^{r_n} \left[O^{\text{cust}}(t) - R^{\text{cust}}(t)\right]dt \tag{3.22}$$

The two interpretations of area A are illustrated in Figure 3.7.

Consider the rectangle indicated by the dashed lines and the axes in Figure 3.7. Let A_R be the area in the rectangle above the restore curve and let A_O be the area in the rectangle above the outage curve. Then $A = A_R - A_O$ where

$$A_R = \left(\sum_{j=1}^{n} c_j^{\text{res}}\right)\Delta r_0 + \sum_{i=2}^{n}\sum_{j=i}^{n} c_j^{\text{res}}\Delta r_{i-1} \tag{3.23}$$

$$A_O = \sum_{k=2}^{n}\sum_{\ell=k}^{n} c_\ell^{\text{out}}\Delta o_{k-1} \tag{3.24}$$

Using the independence of the terms in (3.23), (3.24), the independence of the time differences and the customers out, and $\sum_{i=2}^{n}\sum_{j=i}^{n} 1 = \frac{1}{2}n(n-1)$ gives

$$\overline{A} = \overline{A_R} - \overline{A_O} = n\bar{c}\,\overline{\Delta r_0} + \tfrac{1}{2}n(n-1)\bar{c}(\overline{\Delta r} - \overline{\Delta o}) \tag{3.25}$$

An alternative expression for (3.25) can be obtained using (3.17):

$$\overline{A} = n\bar{c}\overline{D_E} - \tfrac{1}{2}n(n-1)\bar{c}(\overline{\Delta r} + \overline{\Delta o}) \tag{3.26}$$

The terms in (3.25) and (3.26) can be understood by examining the corresponding areas in Figure 3.8. For example, the area $\overline{A_R}$ above the average restore curve is the area of the rectangle with sides $\overline{\Delta r_0}$ and $n\bar{c}$ plus the area of the triangle with sides $(n-1)\overline{\Delta r}$ and $n\bar{c}$. And the area $\overline{A_O}$ above the average outage curve is the area of the triangle with sides $(n-1)\overline{\Delta o}$ and $n\bar{c}$.

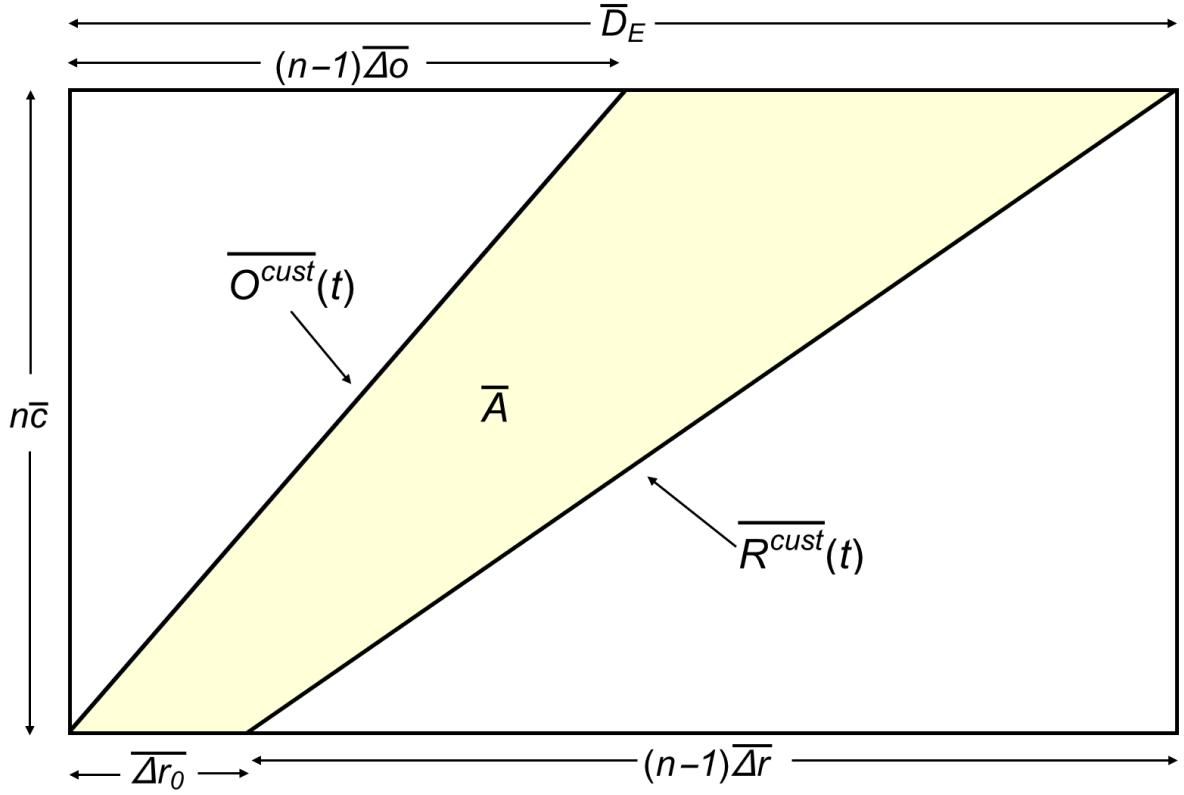

Figure 3.8: Averaged dimensions and customer outage and restoring processes shown to calculate the customer hours $\overline{A}$ in (3.25) and (3.26)

It is useful to describe the outage and restore processes with separate parameters and separate metrics because they respond to different resilience investments. For example, a program of renewing or strengthening components would affect the outage process whereas an increased number of repair crews would affect the restoration process. In more detail, (3.17) shows the effects on the average event duration of reducing the number of outages n (by hardening the infrastructure), reducing $\overline{\Delta r_0}$ (by deploying more inspection crews), and reducing the average time between restores $\overline{\Delta r}$ (by deploying more repair crews). For a larger event, n is larger and deploying more repair crews will have a larger effect because $\overline{\Delta r}$ is multiplied by $n-1$. Formula (3.25) shows the corresponding effects on the customer hours $\overline{A}$. Hardening the upstream system or installing more reclosers can reduce the average customers disconnected per outage $\overline{c}$ and propor-

tionally reduce $\overline{A}$. Reducing the number of outages n or the Δr has an even greater effect for larger events because the second term of (3.25) grows like n^2.

3.4.1 Risk analysis

Customer hours outaged in an event are one of the metrics we can extract from the utility data. Customer hours are an input for SAIDI and ASAI, which are standard metrics utilities use in practice for non-extreme events. The results from the previous section show that customer hours depend on number of component out in an event. Mathematical formulas based on the number of outages (n) were derived from that relationship to approximate resilience metrics such as restoration duration. This section aims to determine how risk measured as the product of customer hours and probability depends on the number of outages n. By way of motivation, note that reducing the number of outages by a certain percentage by hardening can affect this risk.

3.4.2 Obtain and fit empirical distribution of number of components out

We obtain the empirical distribution of the probability of n components out in an event and fit it with a piecewise linear function $p(n)$ on a log-log plot. $p(n)$ joins the parts of two (Zipf-like) linear regions on the log-log plot. The tail part is first estimated using the method of Clauset for discrete power law tails [22]. This gives the value $n = b$ at which the tail starts as well as the exponent of n that determines the slope of the tail on the log-log plot. Then the initial part is estimated using maximum likelihood as summarized below. The linear functions for the initial and tail parts are then combined to give the piecewise linear approximation $p(n)$ to the distribution of n as shown in the following formula and Figure 3.9.

$$p(n) = \begin{cases} \frac{0.410987}{n^{1.36499}} & n = 1, 2, .., 8 \\ \\ \frac{2.66274}{n^{2.26196}} & n > 8 \end{cases} \tag{3.27}$$

Note that although $p(n)$ is a discrete distribution on the positive integers, it might possibly turn out to be convenient for some purposes to regard it as positive integer samples from the continuous piecewise linear function:

$$p(n) = \begin{cases} \frac{0.410987}{n^{1.36499}} & 1 \leq n \leq 8.02993 \\ \frac{2.66274}{n^{2.26196}} & n > 8.02993 \end{cases} \tag{3.28}$$

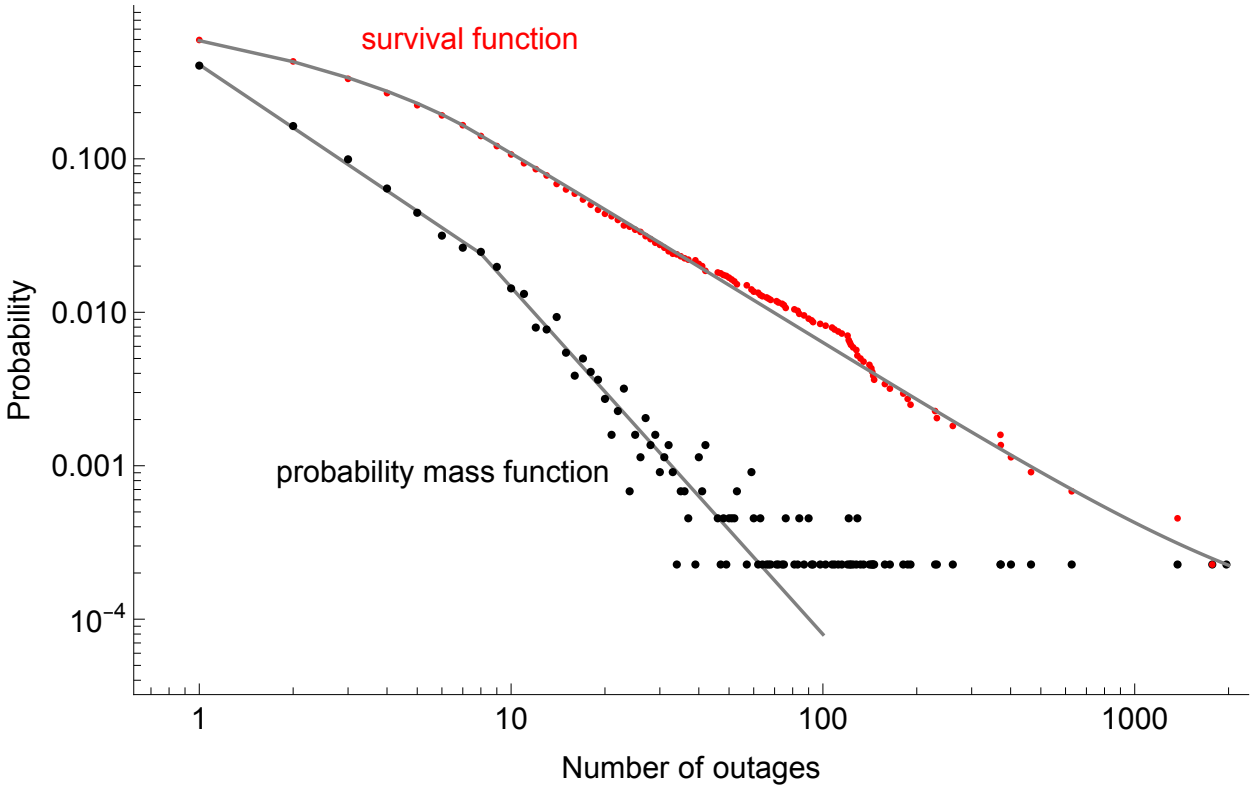

Figure 3.9: Empirical distributions of the number of outages (dots) and their fit with a piecewise linear function $p(n)$ on this log-log plot.

It is straightforward to estimate the maximum likelihood estimate for the initial portion of the distribution as follows: The assumed distribution for the initial portion with slope $-\alpha$ on the log-log plot is

$$p(n) = \frac{n^{-\alpha}}{\sum_{\ell=1}^{b-1} \ell^{-\alpha}}, \qquad n = 1, 2, ..., b-1. \tag{3.29}$$

The likelihood of the data $x_1, x_2, ..., x_N$ that satisfies $1 \leq x_i \leq b-1$ for all $i = 1, 2, ..., N$ is

$$L = \text{likelihood} = (\sum_{\ell=1}^{b-1} \ell^{-\alpha})^{-N} \prod_{i=1}^{N} x_i^{-\alpha} \tag{3.30}$$

Then the log likelihood is

$$\ln L = -N \ln[\sum_{\ell=1}^{b-1} \ell^{-\alpha}] - \alpha \sum_{i=1}^{N} \ln x_i \tag{3.31}$$

and its derivative with respect to α can be set to zero:

$$\frac{1}{N} \frac{d}{d\alpha} \ln L = \frac{\sum_{\ell=1}^{b-1} \ell^{-\alpha} \ln \ell}{\sum_{\ell=1}^{b-1} \ell^{-\alpha}} - \frac{1}{N} \sum_{i=1}^{N} \ln x_i = 0 \tag{3.32}$$

The maximum likelihood value of α can be computed by solving (3.32) numerically.

3.5 Results

This section gives numerical results illustrating the application of the formulas for the statistics of the metrics.

We can evaluate restore duration mean $\overline{D_R}$ and standard deviation $\sigma(D_R)$ for a given number of outages from (3.6) and (3.7). For example, if there are $n = 10$ outages then the restore duration has mean 527 min and standard deviation 211 min. If there are $n = 100$ outages, then the restore duration has mean 3038 min and standard deviation 397 min. If these formulas are to

be used for predicting restoration duration for an incoming storm, then the number of outages n can be predicted by a number of methods as reviewed in the introduction. As well as estimating the mean, it is useful in applying the restore duration to compute its variability with its standard deviation.

The event duration D_E is the restore duration D_R plus the delay until the first restore Δr_0. From (3.8), Δr_0 has mean 132 min and standard deviation 92.4 min. Then we can evaluate event duration mean and standard deviation from (3.17) and (3.18). For example, if there are $n = 10$ outages, then the event duration has mean 660 min and a standard deviation of 230 min. If there are $n = 100$ outages, then the event duration has mean 3171 min and a standard deviation of 408 min.

When an outage event occurs, the primary question the customers want an answer for is "How long will the power be out?". To help determine what should be announced to the public to answer this question, it is useful to estimate an upper bound on the restore duration that will be satisfied with a specified confidence level.

For each given value of number of outages n, the restore duration D_R approximately follows a gamma distribution. We can estimate from (3.6) and (3.7) the mean and standard deviation of D_R and then calculate the gamma distribution with that mean and standard deviation. That is, we estimate the gamma distribution with the method of moments. Then we can easily evaluate the 95th percentile of the gamma distribution. This estimates an upper bound on the restore duration that exceeds the actual restore duration with probability 0.95. The curves in Figure 3.10 show an increasing and initially decelerating increase of mean restore duration as the number of outages increase, and a similar increase in the 95^{th} percentile of restore duration. The estimated best fit function for the empirical restore durations was

$$\overline{\Delta_r} = 7.63765 + 30.7932e^{-0.0514344n} + 33.8291e^{-0.00391306n} \tag{3.33}$$

$$\overline{\sigma(\Delta_r)} = 35.2947 + 43.7128e^{-0.0224135n} \tag{3.34}$$

used to obtain the shape parameter α and scale β for the gamma distribution. The dots in Figure 3.10 are the restore durations for the events in the data; they show how the mean and 95^{th} percentile of restore duration calculated from the estimated gamma distribution summarize the empirical data. The event data becomes sparser as the number of outages increase.

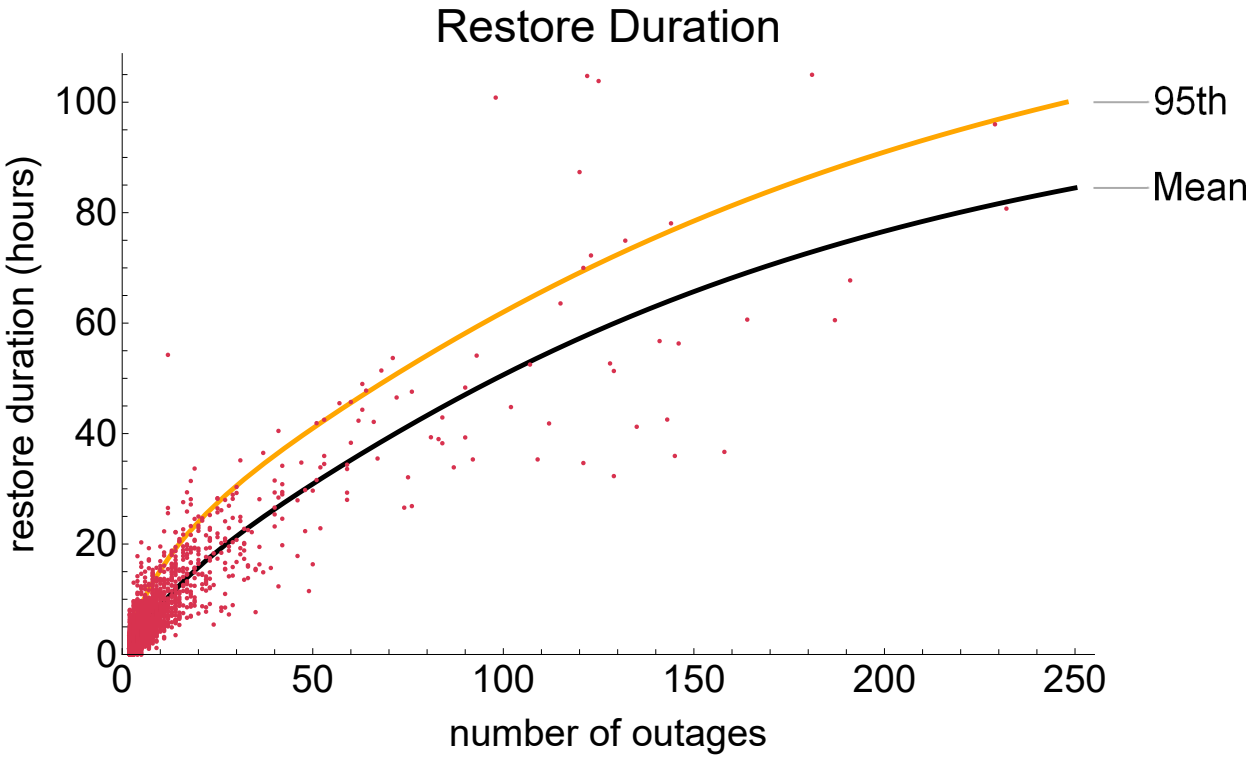

Figure 3.10: Curves show mean and 95th percentile of restore duration D_R versus number of outages. Dots show the restore durations of events in the data.

Figure 3.11 shows the outage rate λ_O and the restore rate λ_R obtained from (3.20) and (3.19) as the number of outages varies. Both rates increase significantly as the number of outages increase. The outage rate results from the interaction of the weather with the grid, and depends on the design margin, age, and maintenance of the grid components.For up to 250 outages, the restore rate is quite close to the outage rate. For more than 50 outages, the restore rate slightly lags the outage rate, showing the extent to which the utility restoring process succeeds in keeping up with the outage process.

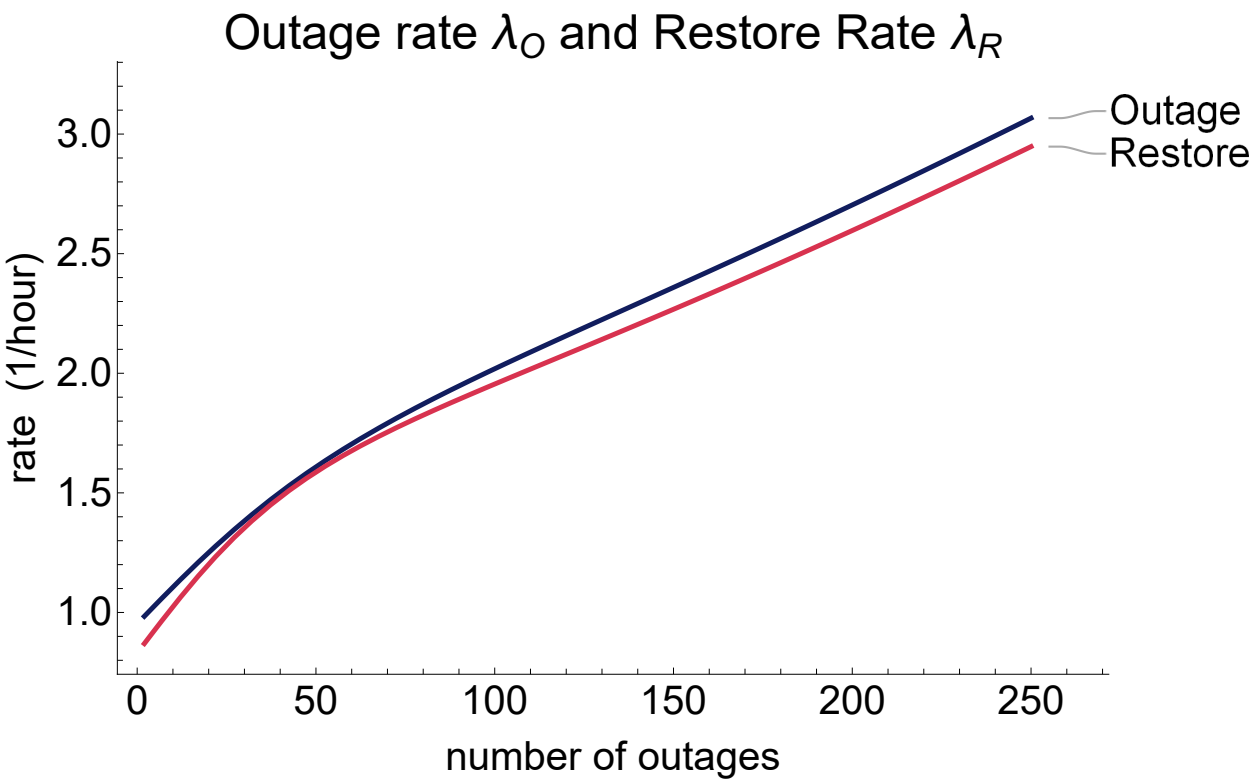

Figure 3.11: Outage rate λ_O and restore rate λ_R versus number of outages.

The curve in Figure 3.12 shows the mean customer hours $\overline{A}$ calculated from (3.25) increasing as a function of the number of outages. The dots in Figure 3.12 show the customer hours A for each event in the data. There is considerable variability in the customer hours in the data for more than 100 outages. Future work could aim to analyze and quantify this variability.

Although sections 3.3.3 and 3.3.4 fit the data for the full range of our data up to 2000 outages, the data for the events with more than 250 outages becomes sparse and more variable. In order to be cautious in our conclusions, we limit all the presented results in this section to events with up to 250 outages. Future work with more data or with more elaborate statistical methods might well extend the range of prediction.

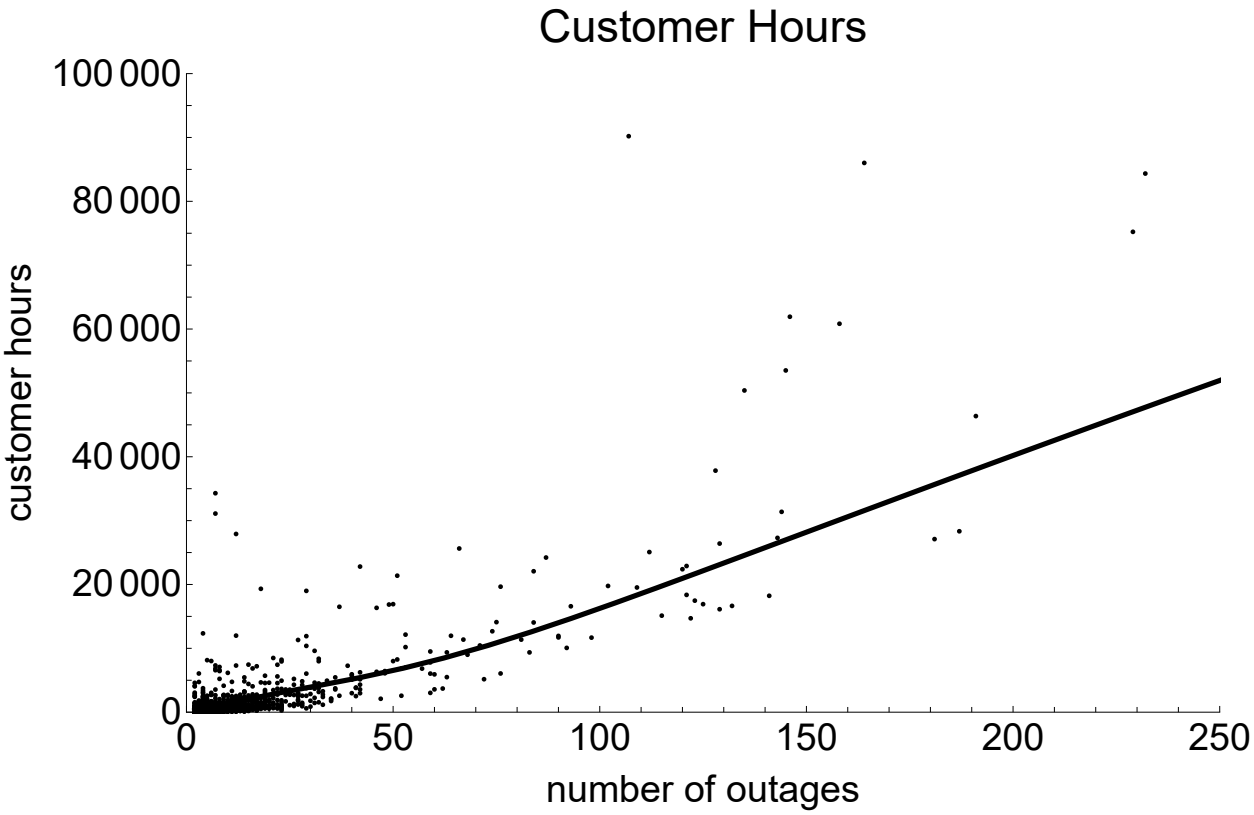

Figure 3.12: Curve shows mean customer hours $\overline{A}$ calculated from (3.25) versus number of outages. Dots show customer hours A of the events in the data.

3.5.1 Risk as a function of number of outages

Let $\overline{A}$ be the average customer hours lost in an event. $\overline{A}$ is also the area under the customer resilience curve for an event. We reproduce the previous formula (3.26) for $\overline{A}$ in terms of the number of outages n in the event from section 1.4.4:

$$\overline{A} = n\overline{c}\,\overline{\Delta r_0} + \tfrac{1}{2}n(n-1)\overline{c}(\overline{\Delta r} - \overline{\Delta o}) \tag{3.35}$$

Evaluating (3.35) for n from 1 to 2000, we obtain the log-log plot of $\overline{A}$ as a function of n as shown in Figure 3.13.

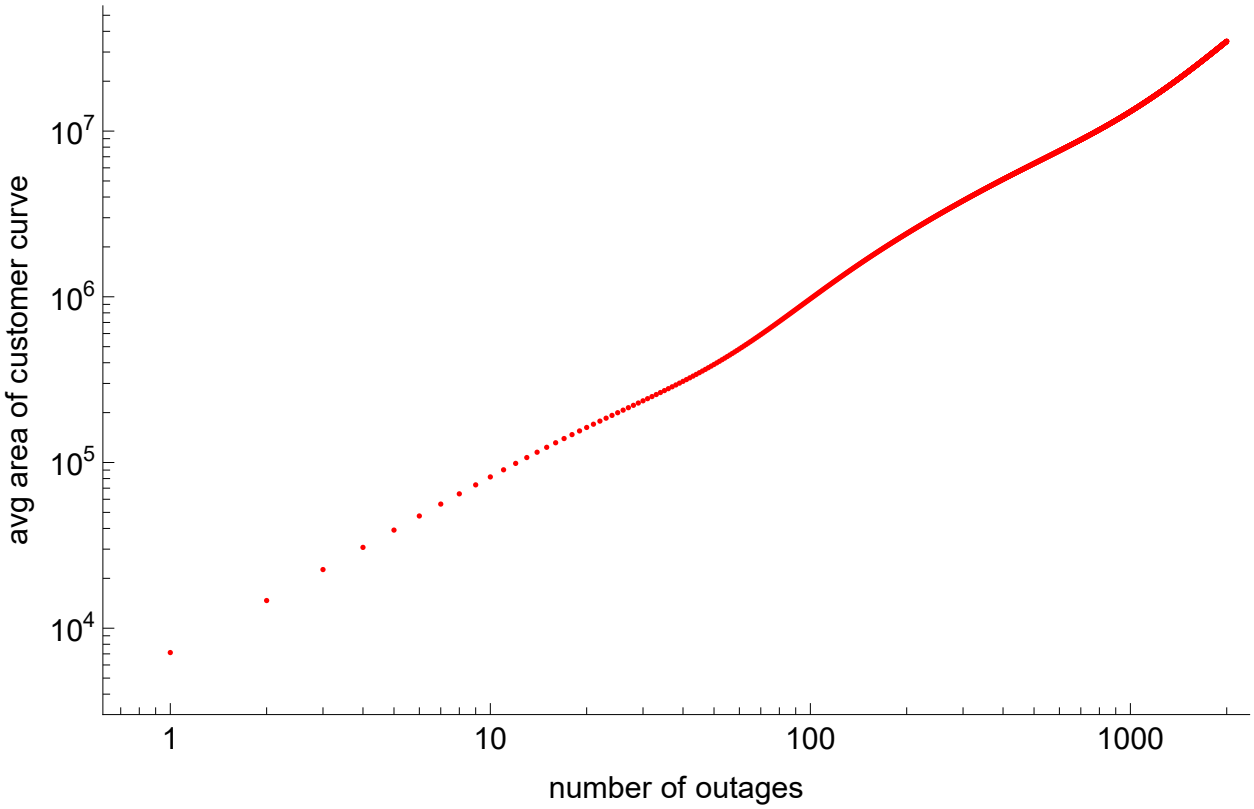

Figure 3.13: Average customer hours lost $\overline{A}$ in an event as a function of number of outages n.

The risk R is calculated by the product of impact (customer hours lost) and probability as

$$R(n) = \overline{A}(n) \times p(n) \tag{3.36}$$

$\overline{A}(n)$ is the customer impact (3.35). $p(n)$ is the probability of n components out in an event obtained in (3.29).

Figure 3.14 shows the risk as a function of n from (3.36). In Figure 3.14 has roughly a power law relationship with the number of outages. We observe that the larger the number of outages the smaller the customer risk.

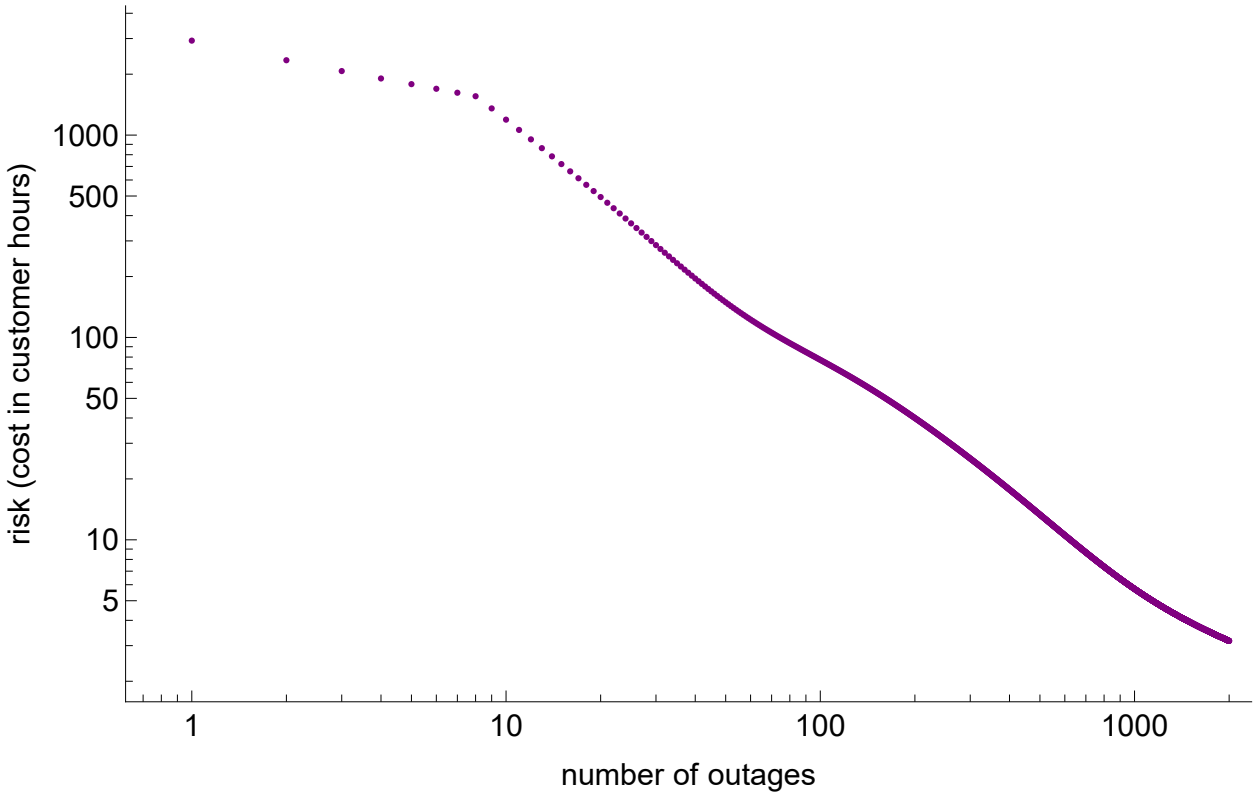

Figure 3.14: Risk as a function of the number of outages on a log-log scale.

For example, for a small n such as 10 outages the risk is very large at roughly 1192.07 customer hours. A similar scenario of a large n such as 1000 outages shows that the risk has relatively small cost at 5.72 customer hours. These two scenarios shows that there is an inverse relationship in the risk between the number of outages and the customer hours.

3.6 Conclusions

We process 5 years of distribution system outage data to extract and study many resilience events in which outages accumulate and are restored. As appropriate for quantifying resilience, we focus only on the resilience events, and do not analyze the frequency of these events or the times between events that are of interest in other kinds of reliability analysis.

It is usual to separate resilience curves for events into successive non-overlapping phases in time such as outage and recovery. However, our distribution utility data shows that outage and

restore processes typically occur together for most of the event. Therefore, instead of using successive phases, we show how resilience curves tracking the number of outaged components or customers can be easily decomposed into outage and restore processes that can occur at the same time. These outage and restore processes describe the same information as the resilience curve, but usefully correspond to different aspects of resilience: the outages are caused by weather interacting with the strength of the components, whereas the restores are done by utility crews and automatic or remote switching. The decomposition of the resilience curve into outage and restore processes is known as the Jordan decomposition in mathematics.

We compute some basic statistics of the outage and restore processes. In particular, we estimate fits for the mean and standard deviations of the times between outages and the times between restores as functions of the number of the outages. The function fitting has the effect of smoothing and interpolating the noisy data. We also estimate the means and standard deviations of the number of customers outaged and of the delay until the restore process starts.

Then, given the predicted number of outages, which is estimated in several ways by previous work [9, 10, 12–15, 23], we obtain formulas for the means of standard resilience metrics, such as restore and event durations, restore and outage rates, the customer hours lost and its risk as a cost. These formulas for standard resilience metrics usefully quantify the resilience processes and show how the metrics depend on the number of outages, the delay before restoration starts, the average time between restores, and the average number of customers disconnected per outage. The outage rate quantifies the overall grid fragility under weather stress and the restore rate quantifies the overall performance of utility crews. This quantification can help inform investments that improve these metrics. We also estimate the standard deviations of the restore duration and the event duration. This leads to estimates of probable upper bounds of restore durations. These credible upper bounds based on past performance should be useful to utilities for informing customers about their outage duration with confidence. The utility data available to us limited our resilience processes to counting outages of components or customers. If available, quantities such as power outaged could be similarly analyzed to obtain useful metrics.

Our approach models the outage and restore processes directly from utility data and avoids the complexities of modeling individual component restoration or repair times and their order of completion. That is, since the utility data itself incorporates the detailed complexities of resilience, we can give a high-level description and quantification of resilience. This high-level data-driven approach is very much a useful complement to the detailed modeling of the restoration processes by other authors.

The average customer risk for a given size of event can be evaluated as that product of the average customers lost and the probability of event size. For the utility data examined, the average customer risk decreases as the event size increases.

In summary, we extract and separate the outage and restore processes from distribution utility outage data in a new way, estimate the statistics of times between successive restores or outages, and then show how standard resilience metrics can be derived from these statistics. The overall effect is to compute some useful resilience metrics from practical utility data.

3.7 References

[1] N. K. Carrington, I. Dobson, and Z. Wang, "Extracting resilience metrics from distribution utility data using outage and restore process statistics," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5814–5823, 2021.

[2] C. Nan and G. Sansavini, "A quantitative method for assessing resilience of interdependent infrastructures," *Reliability Engineering & System Safety*, vol. 157, pp. 35–53, 2017.

[3] M. Panteli, D. N. Trakas, P. Mancarella, and N. D. Hatziargyriou, "Power systems resilience assessment: hardening and smart operational enhancement strategies," *Proceedings IEEE*, vol. 105, no. 7, pp. 1202–1213, 2017.

[4] M. Ouyang, L. Dueñas-Osorio, and X. Min, "A three-stage resilience analysis framework for urban infrastructure systems," *Structural safety*, vol. 36, pp. 23–31, 2012.

[5] N. Yodo and P. Wang, "Resilience modeling and quantification for engineered systems using bayesian networks," *Journal of Mechanical Design*, vol. 138, no. 3, p. 031404, 2016.

[6] N. K. Carrington, S. Ma, I. Dobson, and Z. Wang, "Extracting resilience statistics from utility data in distribution grids," in *2020 IEEE Power Energy Society General Meeting (PESGM)*, 2020, pp. 1–5.

[7] C. J. Zapata, S. C. Silva, H. I. Gonzalez, O. L. Burbano, and J. A. Hernandez, "Modeling the repair process of a power distribution system," in *2008 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America*, 2008, pp. 1–7.

[8] Y. Wei, C. Ji, F. Galvan, S. Couvillon, G. Orellana, and J. Momoh, "Non-stationary random process for large-scale failure and recovery of power distribution," *Applied Mathematics*, 2016.

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

[10] H. Liu, R. A. Davidson, D. V. Rosowsky, and J. R. Stedinger, "Negative binomial regression of electric power outages in hurricanes," *Journal of infrastructure systems*, vol. 11, no. 4, pp. 258–267, 2005.

[11] S. Kancherla and I. Dobson, "Heavy-tailed transmission line restoration times observed in utility data," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1145–1147, Jan 2018.

[12] D. Zhu, D. Cheng, R. P. Broadwater, and C. Scirbona, "Storm modeling for prediction of power distribution system outages," *Electric Power Systems Research*, vol. 77, no. 8, pp. 973–979, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378779606001969

[13] Y. Zhou, A. Pahwa, and S. . Yang, "Modeling weather-related failures of overhead distribution lines," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1683–1690, 2006.

[14] K. Alvehag and L. Soder, "A reliability model for distribution systems incorporating seasonal variations in severe weather," *IEEE Transactions on Power Delivery*, vol. 26, no. 2, pp. 910–919, 2011.

[15] H. Li, L. A. Treinish, and J. R. M. Hosking, "A statistical model for risk management of electric outage forecasts," *IBM Journal of Research and Development*, vol. 54, no. 3, pp. 8:1–8:11, 2010.

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

[17] M.-Y. Chow, L. S. Taylor, and M.-S. Chow, "Time of outage restoration analysis in distribution systems," *IEEE Transactions on Power Delivery*, vol. 11, no. 3, pp. 1652–1658, 1996.

[18] H. Liu, R. A. Davidson, and T. V. Apanasovich, "Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms," *Reliability Engineering & System Safety*, vol. 93, no. 6, pp. 897–912, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0951832007001305

[19] "Jordan decomposition (of a function)." [Online]. Available: http://encyclopediaofmath.org/index.php?title=Jordan_decomposition_(of_a_function)&oldid=29165

[20] T. Jafarikhah and K. Weihrauch, "Computable Jordan decomposition of linear continuous functionals on $c[0;1]$," *Logical Methods in Computer Science*, vol. 10, 2014.

[21] A. E. Taylor, *General Theory of Functions and Integration.* Dover Publications, 1985.

[22] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[23] P. Kankanala, S. Das, and A. Pahwa, "Adaboost$^+$: An ensemble learning approach for estimating weather-related outages in distribution systems," *IEEE Transactions on Power Systems*, vol. 29, no. 1, pp. 359–367, 2014.

CHAPTER 4. WEATHER IMPACT ON RESILIENCE

Nichelle'Le K. Carrington, Ian Dobson, and Zhaoyu Wang, Department of Electrical and Computer Engineering Iowa State University, Ames, Iowa, USA

Modified from a manuscript to be submitted to a future *IEEE Power & Energy Society General Meeting*

4.1 Abstract

We examined the outage cause codes of the resilience events found using the event definition in chapter 3. Tree limb and weather reasons were determined to be the cause for majority of events in relation to event size. An exploratory survey of wind speed and its relation to event size was performed. The wind speed data from NOAA LCD was compressed and combined with the utility data. The results show how the wind speed increases as the resilience event size increases.

4.2 Overview

According to the Congressional Research Service's 2012 report on weather-related power outages and electric system resilience, tree limbs and high winds from seasonal storms are the main causes of prolonged outages [1]. In the interest of effective system hardening, detailed distribution outage data containing geographical and outage cause information can help understand the cause of outage events. Moreover, weather station data can provide detail about atmospheric conditions during outages. Since utility data sets are typically heterogeneous using the causes codes associated with the outages makes it possible to gain more insight into the reasons for outages. This chapter focuses on determining which cause codes contribute the most to outages and resilience events.

4.3 Individual Outage Causes

In the distribution utility outage data described in 1.3.1, 64 different cause codes record the reason for each outage. The cause code "Unknown " is assigned to a component outage if the cause is missing or not reported. The 64 original cause codes categorize into seven group causes as follows: "Weather," "Tree," "Treelimb," "Animal," "Human," "Other," and "Equipment." There are 30 094 outages in the distribution utility data and the original cause codes were replaced with the associated group cause. The outages are broken down by their group as follows: 10882 are "Treelimb," 4753 are "Equipment," 3595 are "Animal," 1228 are "Tree," 3784 are "Other," 840 are "Human," and 5012 are "Weather." The "Treelimb" cause code group, an outage occurs where a tree limb (branch) is either inside or outside the clearance area of the component, and is different from the "Tree" cause code group which occurs when parts of a tree cause damage to a component from trimming, a tree dropping on a component, or a utility worker error caused by a tree. Figure 4.1 shows the breakdown of the event outages by the group outage causes.

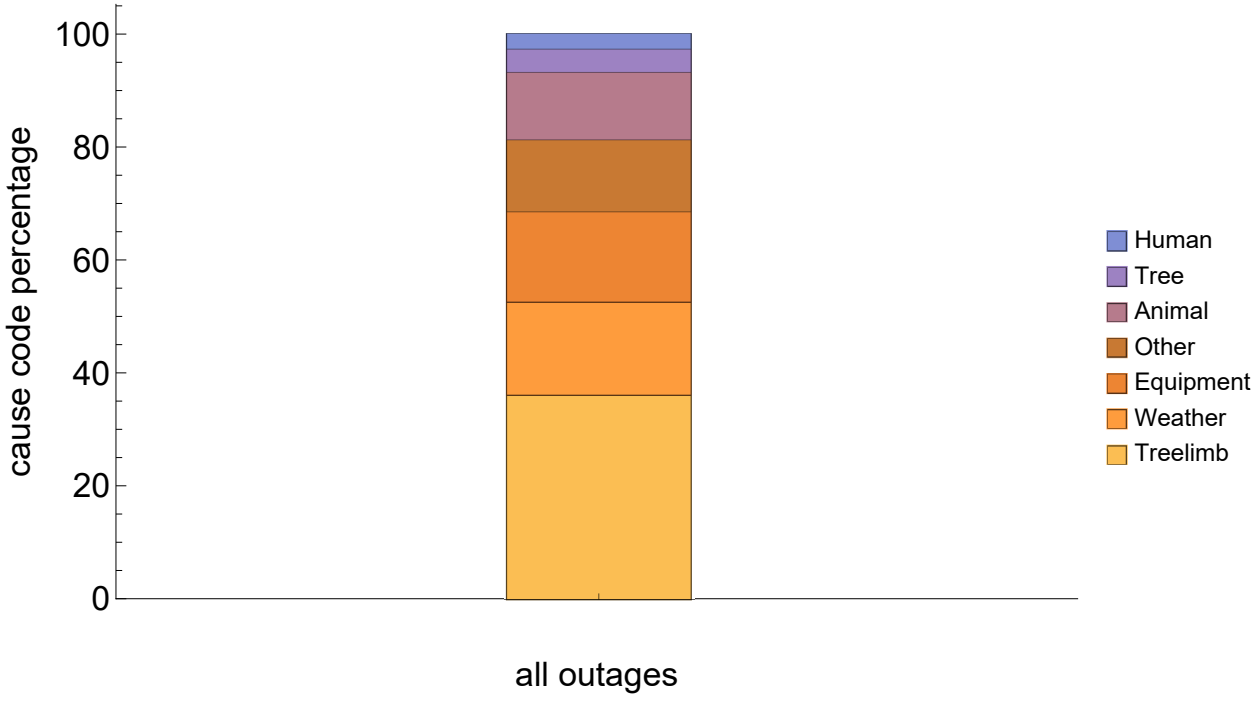

Figure 4.1: Breakdown of cause groups for all outages

According to the data, tree limb damage and weather are the two leading causes of outages. Additionally, we examined the composition of outages of small, medium and large events in Figure 4.2. Tree limbs are the largest cause for outages in all sizes of event. For small ($2 \leq n \leq 29$) and medium ($30 \leq n \leq 100$) size events, equipment-related outage causes are the second largest cause of outages. In large ($n > 100$) size events, weather-related outage causes are the second largest cause of outages.

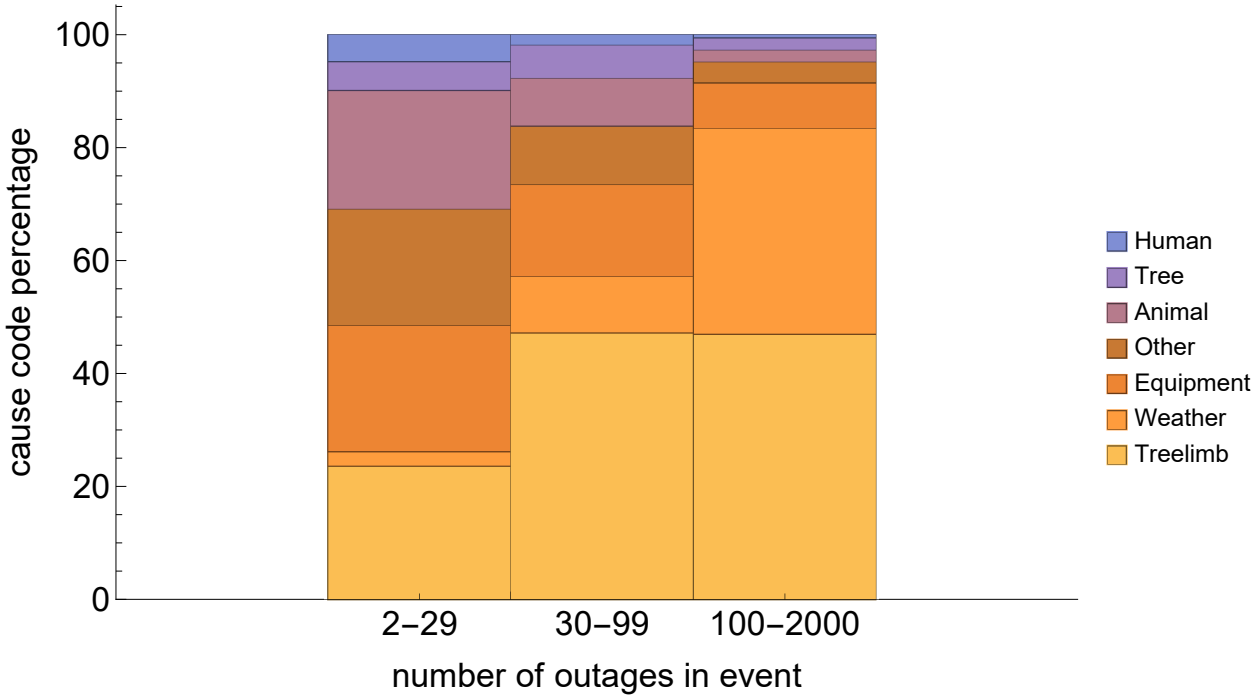

Figure 4.2: Breakdown of cause groups for outages in small, medium, and large events.

Based on this breakdown of the outage causes, there is a strong presence of tree limb outages and significant amounts of weather-related outages. Although this is the case, it does not mean that the majority of outages are caused by each of these factors. However, for large events, outages groups "Tree limb" and "Weather" together account for 52% of outages.

4.4 Majority Causes of Outages in Events

The outages in a given resilience event often have several different cause codes. Instead of analyzing causes of each individual outage, we can analyze the cause of an event by its majority cause. The majority cause code for an event is the cause with the highest percentage of all outage causes in that event. A random choice is used when there is a tie between two or more causes to determine the majority cause.

Figure 4.3 breaks down the majority causes of all events. According to Figure 4.3, the most common group event causes are almost equally divided between "Treelimb", "Equipment", "Animal", and "Other", with slightly more "Treelimb" events. Few events had the majority cause group of "Weather".

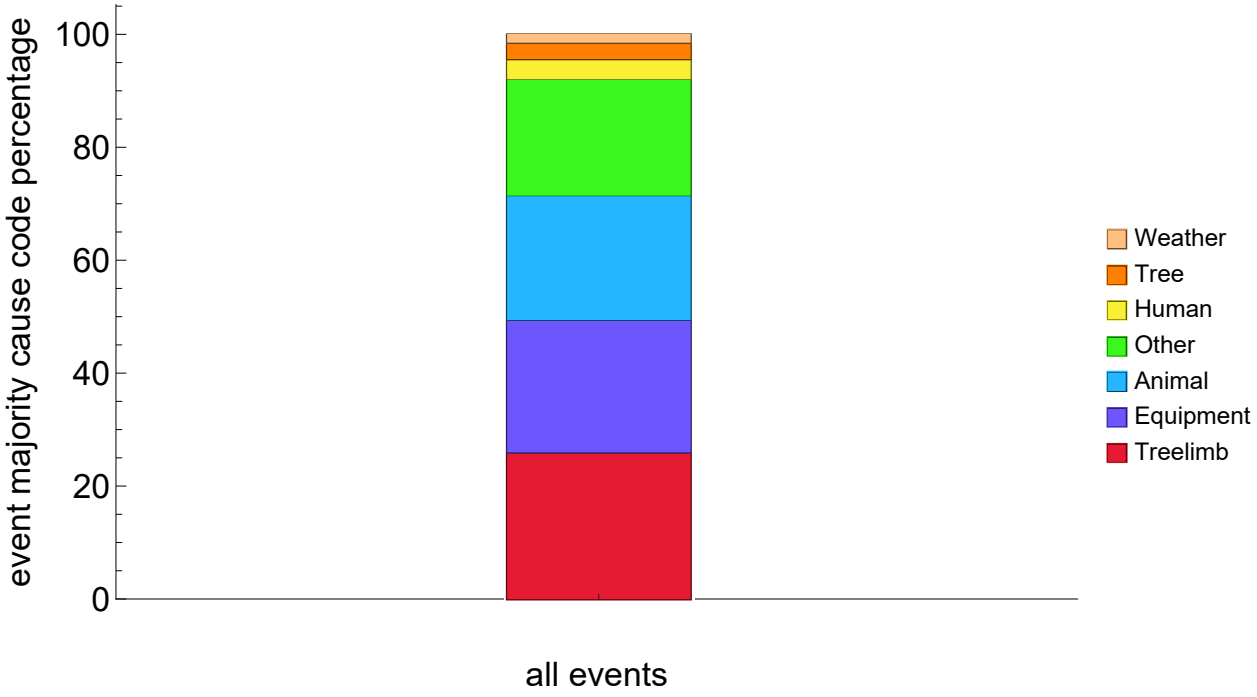

Figure 4.3: Breakdown of majority cause groups for all events.

However, the majority cause of events strongly depends on event size as shown in Figure 4.4. The majority group causes of medium and large events are dominated by "Treelimb" causes. For medium size events, outages groups "Treelimb" and "Weather" together account for 89% of out-

ages. For large size events, outages groups tree limb and weather together account for 97% of outages. We can conclude that while small events have a range of different causes, medium and large events show some homogeneity in cause. Since "Treelimb" causes can largely be associated with wind, this shows that medium and large events can be mainly attributed to weather.

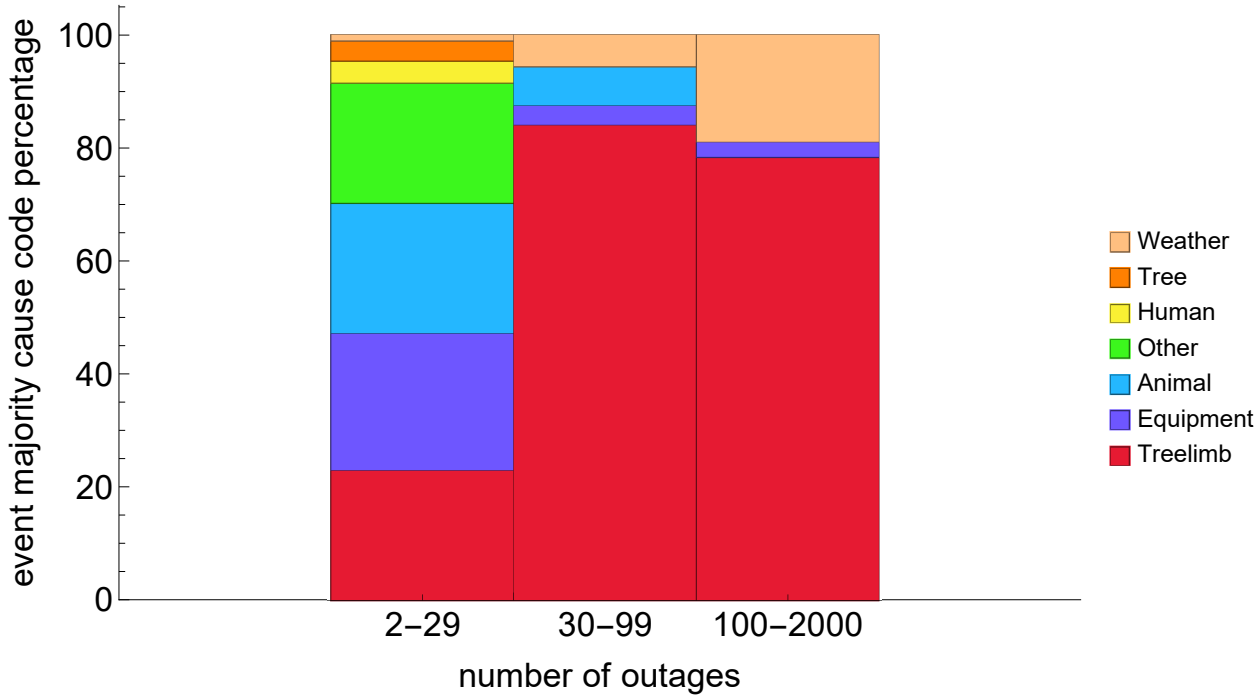

Figure 4.4: Breakdown of majority cause groups for small, medium, and large events.

4.5 Outage and Weather Data Processing

Detailed distribution utility outage data were cleaned and processed for use in chapter 2. Longitude and latitude information provided by the utility outage dataset helped identify the components' regions. It also shows what weather stations are within each region that record weather measurements as shown in Figure 4.5. Climate information, such as temperature, precipitation, and wind speed, was collected at the same time as utility outage data by the National Oceanic and Atmospheric Administration (NOAA). This work focuses on the sub-region of the utility footprint shown in Figure 4.5.

Figure 4.5: The gray points are markers for the location of components in the network that were listed in the outage data. The pink dots indicate a weather station.

4.5.1 Wind Speed Data

The Local Climatological Data (LCD) is a data set from NOAA that contains wind speed and other wind-related measurements from on-land locations[2]. LCD records date back to 2005 and are only available for 1000 US stations and include daily extremes, average temperatures, precipitation, wind speed, and wind direction. I obtained the stations' locations and LCD data for Region B from the NOAA website. In the LCD, the hourly wind speed is measured in miles per hour and is the wind speed at the time of recording.

4.5.2 Data Compression

The data compression ans coordination of outage and wind speed data is a straight forward process. The date and hour of and each outage that occurred in Region B is extracted from the detailed utility outage data. The total number of outages is calculated for each date and hour and recorded for that timestamp, shown in Table 4.1.

Table 4.1: Hourly Utility Outage Data

Outage Date Hour	Total Number of Outages
2/18/2011 21:00	1
2/18/2011 23:00	4
2/19/2011 00:00	2
2/19/2011 01:00	2
2/19/2011 01:00	2
⋮	⋮

In Table 4.2, the date, hour and wind speed are extracted from the LCD data from NOAA of the weather stations (pink dots) in Figure 4.5. Missing wind speed values are substituted with 0.

Table 4.2: Hourly Wind Speed

Date Hour	Hourly Wind Speed (mph)
1/1/2011 00:00	3
1/1/2011 01:00	2
1/1/2011 02:00	4
1/1/2011 03:00	0
⋮	⋮

The wind speed and the total number of outages from Tables 4.1 and 4.2 are obtained for the outages in each event.

4.6 Distribution of Wind speed as event size increases

From Section 4.5 we are able to find how the mean wind speed depends on event size as shown in Figure 4.6.

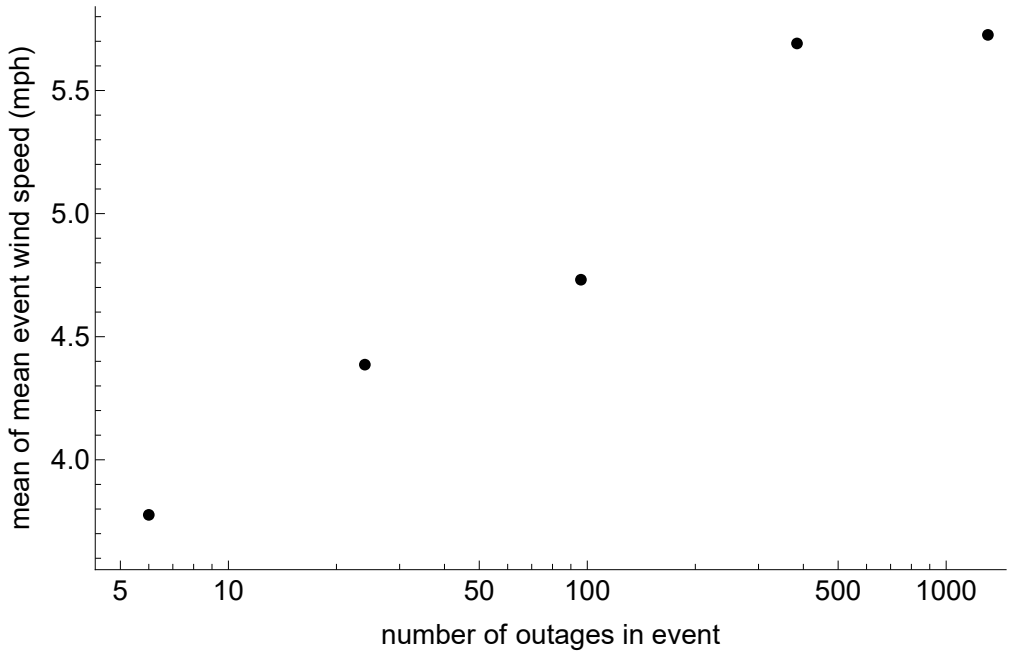

Figure 4.6: Mean event wind speed as a function of event size.

Figure 4.6 shows that as event size increases the average wind speed increases as well. The same can be said for the mean maximum mean event wind speed in Figure 4.7.

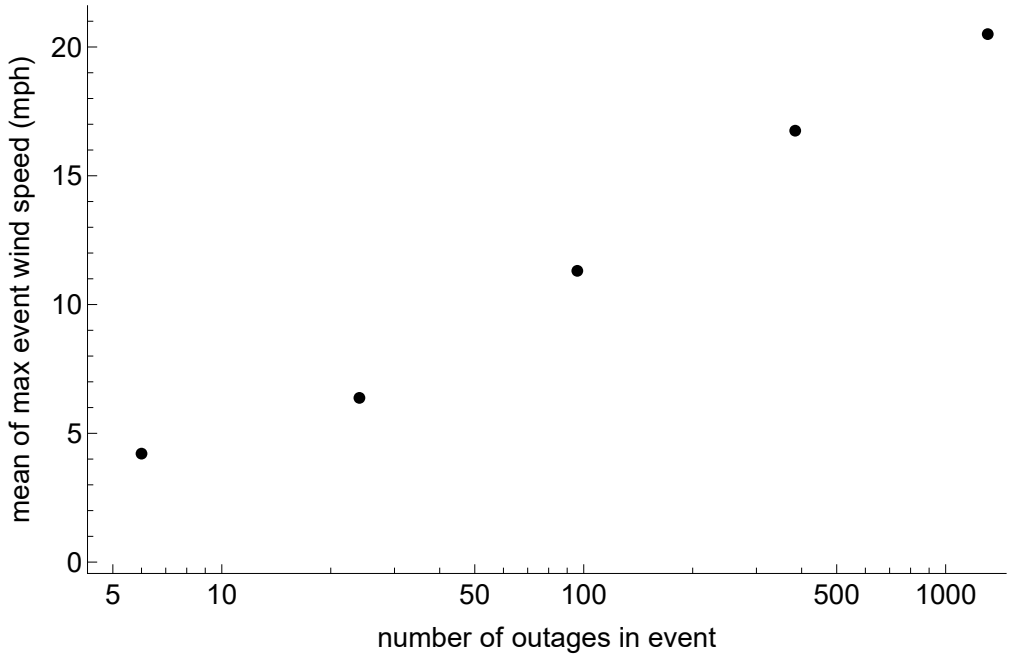

Figure 4.7: Mean maximum event wind speed of outages as a function of event size.

4.7 Conclusion

The work in this chapter examined the outage cause codes for the data set described in section 1.3.1 in chapter 1. In general it a was determined that "Treelimb" was the primary cause for outages and the majority cause for events. In this data set, "Weather" and "Treelimb" related outages accounted for over 80% of the majority cause of outages for the medium and large events. These findings showed that as the event size increases the cause for outages in this data set become more homogeneous.

An exploratory survey of wind speed and its relation to event size was performed. Wind speed is a common metric used in models and simulations for system resiliency. The wind speed data from NOAA LCD was compressed and combined with the utility data. The results show how the wind speed increases as the resilience event size increases.

4.8 References

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

[2] National Oceanic and Atmospheric Administration (NOAA), "Noaa national centers for environmental information local climatological data." [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/datatools/lcd

CHAPTER 5. EXPLORING CASCADING OUTAGES AND WEATHER VIA PROCESSING HISTORIC DATA

Ian Dobson, Nichelle'Le K. Carrington, Kai Zhou and Zhaoyu Wang,
Department of Electrical and Computer Engineering Iowa State University, Ames, Iowa, USA
Benjamin A. Carreras, BACV Solutions, Oak Ridge, TN, USA
José M. Reynolds-Barredo, Departamento de Física,Universidad Carlos III, Madrid, Spain
Modified from a manuscript published in the 51^{st} Hawaii International Conference on System Sciences (HICSS-51) [1]

5.1 Abstract

We describe some bulk statistics of historical initial line outages and the implications for forming contingency lists and understanding which initial outages are likely to lead to further cascading. We use historical outage data to estimate the effect of weather on cascading via cause codes and via NOAA storm data. Bad weather significantly increases outage rates and interacts with cascading effects, and should be accounted for in cascading models and simulations. There are very good prospects for improving data processing and models for the bulk statistics of historical outage data so that cascading can be better understood and quantified.

5.2 Overview

A cascade of related outages weakens or degrades a transmission system in a progressive fashion [2]. In spite of careful design and operation, there can be cascading outages large enough to cause load shedding and blackouts on our power transmission system. Cascades of blackouts of this magnitude are relatively rare, but pose a substantial risk [3–6].

In general, transmission blackouts spread through cascading, and many factors contribute to initial outages or their subsequent propagation. To assess and mitigate cascading outages, a variety of models, approximations, simulations, and procedures have been developed [2, 7]. Validating these efforts with historical data [8–10] can assist in evaluating and improving them. Utility companies collect outage data much more systematically and automatically now, but extracting and processing useful information from the data remains a challenge.

In this chapter, we report on some bulk statistical processing of 14 years of transmission line outage data from a large North American utility. This processing is to describe initial line outages and to start to explore the effect of weather on cascading. Our data-driven analysis of the effect of weather on the bulk statistics of cascading and aspects of our bulk statistical analysis of initial line outages are novel. A map of cascading across counties is easily obtained using storm data and line outage data.

As an alternative to working directly with data as in this chapter, one can make simulation models that use or are tuned to typical parameter values. This approach has been taken by several authors to propose weather effects models in cascading simulations [11–14].

Historical data processing has many advantages, such as avoiding modeling assumptions and providing a very favorable grounding. However, it is critical to note that the grid evolves over a 14-year period and statistical analysis of historical cascades necessarily describes the risk cascade over time.

5.3 Data description and processing

5.3.1 Transmission Outage Data

The transmission line outage data in this chapter consists of 42 561 automatic and planned line outages recorded by a North American utility over 14 years starting in January 1999 and ending in December 2013 [15]. The automatic line outages are identified within the dataset. This data is standard and routinely collected by North American utilities. This data is reported in NERC's Transmission Availability Data System (TADS) [16, 17] and is also collected in other

countries. The information in the data includes the outage start time (to the nearest minute), names of the buses at both ends of the line, and the dispatcher cause code.

5.3.2 NOAA Storm Data

The National Oceanic and Atmospheric Administration (NOAA) Storm Events Database is a collection of the occurrence of storm events and other significant weather phenomena recorded by NOAA's National Weather Service from 1950 to August 2021 [18]. The NOAA storm data includes the event type, event start and end time, and the location within the state by county or zone. In this chapter, the period of the NOAA storm data used is between January 1999 to December 2013.

5.3.3 Data Processing

The work in [19] formed a network model from the line outages using the sending and receiving bus names to identify connections in the grid. The network model is a connected network and contains 614 lines and 361 buses. Of the 42 561 automatic and planned line outages in the data, 10 942 are automatic. The analysis of the cascades in this chapter focuses on automatic outages only. Each cascade begins with initial outages in the first generation, followed by additional outages categorized into following generations until the cascade comes to an end [20].

The initial stage in processing line failures is to arrange them into distinct cascades and then group the outages in close succession within each cascade into generations. The outages are grouped into cascades and generations within each cascade using [21, 22]. The technique is summarized here, with the specifics found in [22]. The technique uses the gap in start time between successive to categorize the outages. If there is a one-hour or longer break between outages, the outage following the gap starts a new cascade.If successive outages within a cascade have a gap of more than one minute, the outage following the interval initiates a new generation of the cascade.The order of outages within a generation is sometimes impossible to discern since outage times are only available to the nearest minute.

Because operator actions are usually completed within one hour and fast transients and protection actions such as automatic reclosing are usually completed within one minute, this simple method of defining cascades and generations of outages appears to be effective and has gap thresholds consistent with power system time scales. [22] investigates the resilience of cascade propagation when these gap thresholds are varied.

5.4 Effect of weather and other influences via cause codes

A dispatcher's outage cause code allows the classification of the cascades of outages into two categories: weather-related and non-weather-related ones. A cascade of outages is considered weather-related when an outage in the cascade has at least one of the cause codes"Weather," "Lightning," "Galloping Conductors," "Ice," "Wind," or"Tree blown. The field cause code does not factor into this analysis.

Table 5.1: Some general dependencies of initial outages and average propagation

equivalent annual cascade rate	propagation λ	N = number of outages in cascade $P[N>1]$	$P[N>5]$	$P[N>10]$	CAUSE
478	0.28	0.26	0.027	0.007	ALL OUTAGES
101	0.55	0.51	0.096	0.028	WEATHER
377	0.13	0.19	0.009	0.002	NOT WEATHER
588	0.31	0.29	0.04	0.009	SUMMER MONTHS
423	0.25	0.24	0.02	0.006	NOT SUMMER MONTHS
486	0.36	0.34	0.05	0.010	PEAK HOURS
475	0.25	0.24	0.02	0.007	NOT PEAK HOURS

The weather-dependent annual cascade rate, average propagation rate, and cascade size distribution can be viewed in Table 5.1 and Figure 5.1. Based on Table 5.1, only 21% (101/478) of the cascades are weather-related. Hence, less than 21 percent of the initial outages were caused by a cascade of weather events. In Figure 5.1, the distribution of initial outages is the same for weather-related as well as non-weather-related cascades. The propagation of weather-related outages has, however, been greatly increased in Table 5.1 from 0.13 (non-weather related) to 0.55 (weather-related). This is evident with Figure 5.1 in terms of the distribution of the outages in a

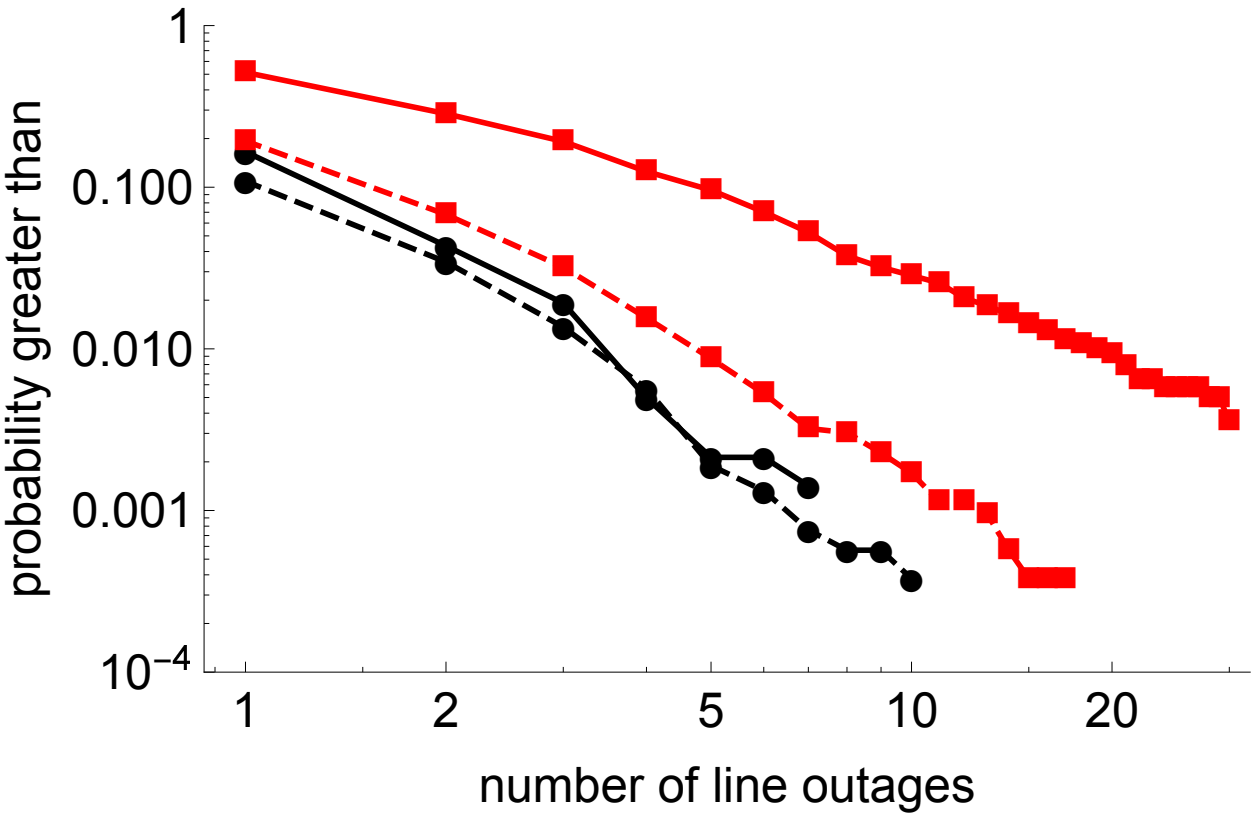

Figure 5.1: Probability distributions of initial (black circles) and cascaded (red squares) outages with weather (solid line) and no weather (dashed line). Weather is determined by cause code.

cascade after cascading. Weather-related cascades make up a small portion of cascades, but they multiply substantially to form larger cascades.

However, for the subset of weather-related outages, the same contention isn't valid because there is a higher rate of independent outages during bad weather. While the methods of Section 5.5 do not provide conclusive results, the results of Section 5.5support this conclusion. Additionally, traditional risk analysis does show that independent outages are much more common during bad weather [23, 24]. As a result, the validity of cascading processing applied to weather-related outages is called into question. Outages that occur due to network interactions are interpreted as dependent outages. Nonetheless, if the concern is simply the number of subsequent outages during one hour without regard to the cause, the method may have some validity for

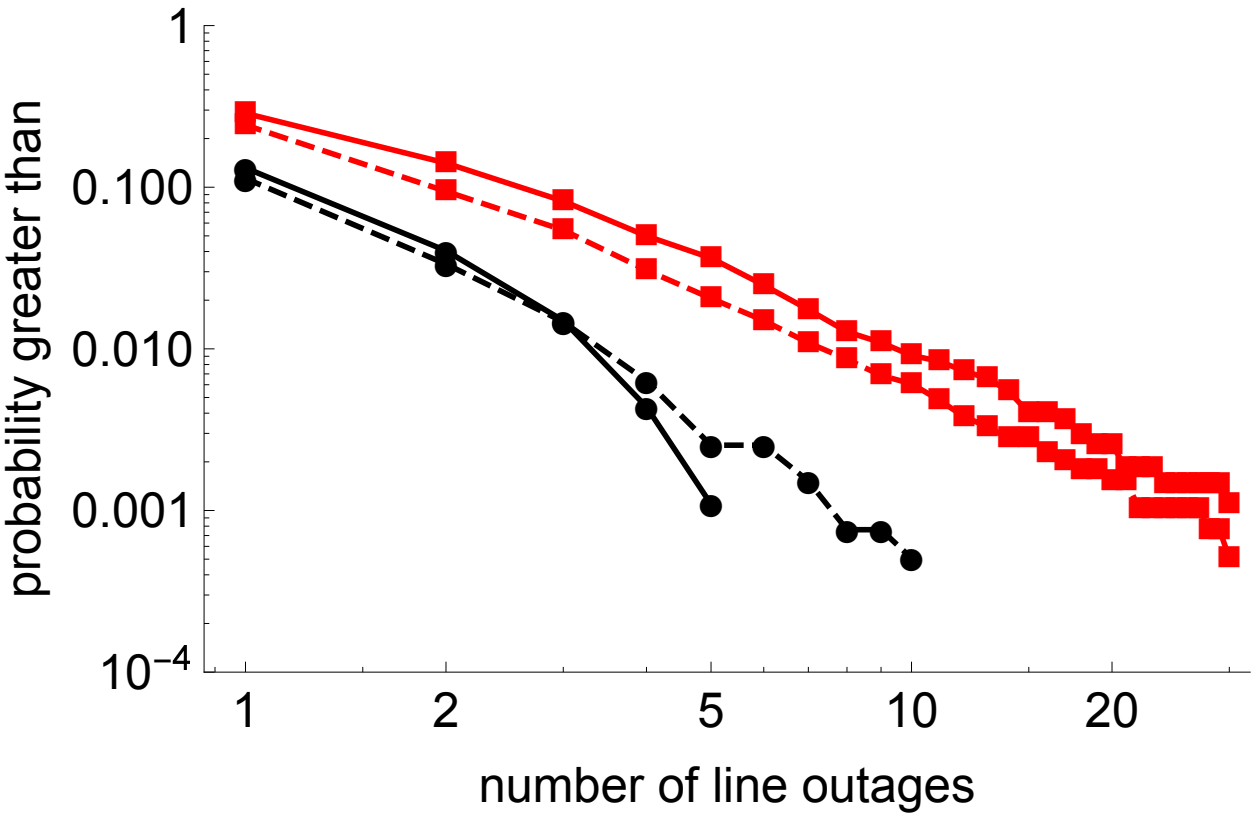

Figure 5.2: Probability distributions of initial (black circles) and cascaded (red squares) outages in summer months (solid line) and remainder of year (dashed line).

weather-related outages. For example, this might be relevant when the operator has to handle the total amount of outages, regardless of cause, within an hour.

Cascades can be classified by month and time of day, i.e., those occurring during the summer peak months of June, July, August, September, as well as those occurring outside these peak hours. Table 5.1 and Figures 5.2 and 5.3 show the impact of the summer months and the peak hours. The equivalent annual rate in Table 5.1 represents the rate if conditions such as summer months were applicable all year long.) Outages in the summer months of June, July, August, September have a modestly increased propagation from 0.25 (not summer) to 0.31 (summer). In the peak hours between 3 pm and 8 pm, outages have increased propagation from 0.25 (not peak hours) to 0.36 (peak hours). Note that cascades are also affected by initial outages. Indeed, summer months show 40% more initial outages and 39% more cascades. On average, there is a

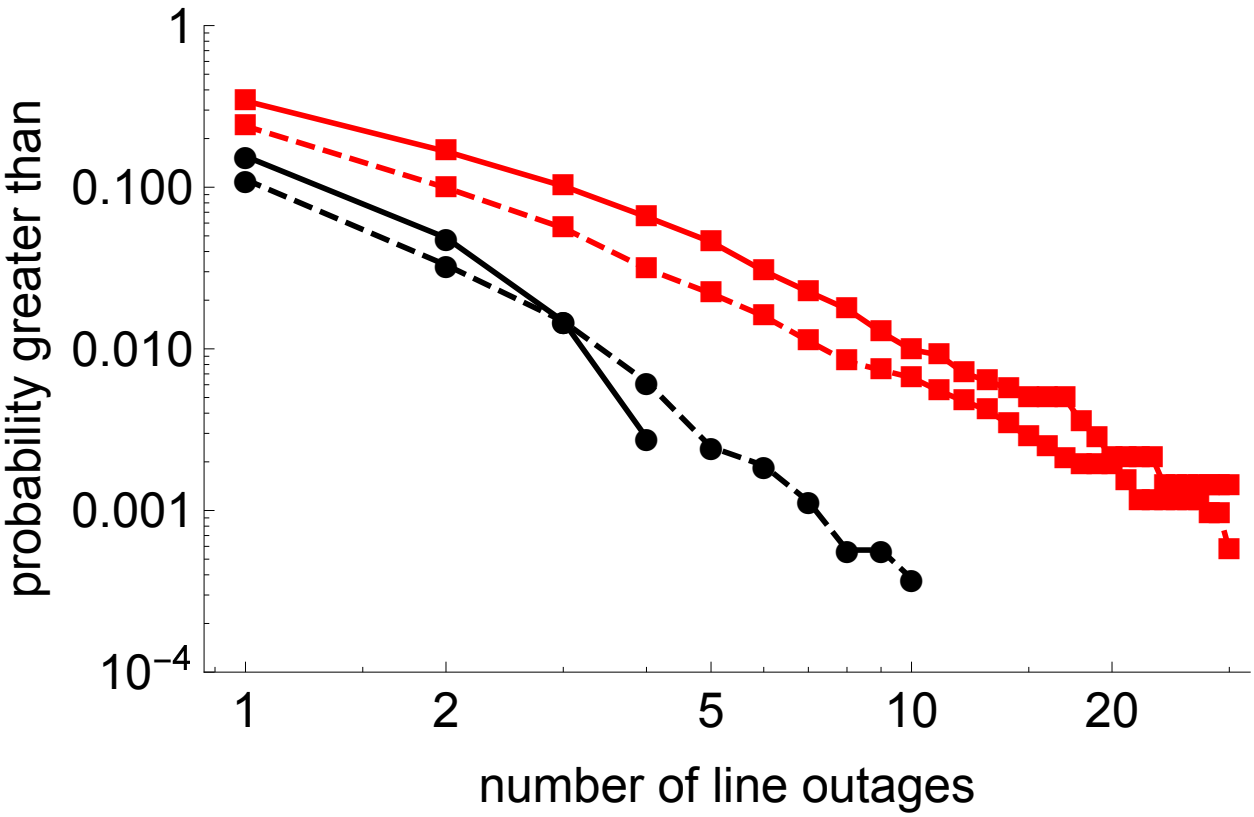

Figure 5.3: Probability distributions of initial (black circles) and cascaded (red squares) outages at peak hours (solid line) and off-peak hours (dashed line).

moderate increase in cascade propagation during peak hours and only a small increase in propagation, but an increased rate of initial outages in the summer. Weather effects, however, are greater than either of these factors.

5.5 Effect of weather via NOAA weather data

Analysis of weather effects on outage cause codes has the disadvantage that cause codes cannot describe the weather when there is no outage. To that end, key quantities such as the outage rate in bad weather cannot be determined from cause code analysis.

In addition, the outage cause codes are manually entered, rely on a subjective best judgment about conditions and classifications, and include a significant proportion (22% of the dispatcher

outage cause codes) of causes "Unknown". Coordination of outage data with weather records is one approach to solving these problems.

The National Oceanic and Atmospheric Administration (NOAA) Storm Events Database is a collection of the occurrence of storm events and other significant weather phenomena recorded by NOAA's National Weather Service from 1950 to present [18]. The NOAA historical storm data records for 1999 to 2013 were obtained for analyzing the storm weather effects influencing our outage data. The NOAA storm data includes the event type, event start and end time, and the location within the state by county or zone. The storm event types that we choose to define as a storm for analyzing the power grid are "Blizzard", "Freezing Fog", "Hail", "Heavy Rain", "Heavy Snow", "High Wind", "Ice Storm", "Lightning", "Sleet", "Strong Wind", "Thunderstorm Wind", "Tornado", "Winter Storm", "Winter Weather".

The bus outages are associated with the storm data by mapping the buses to the county they are located in, and then describing the zones by the counties they intersect. A line is defined as a county if either its sending or receiving end bus is in that county. Zones are defined as lines that include a county that the line is in. This associates each line with a set of counties. In some cases, the set may contain only one county. It is straightforward to count the number of storm outages of a particular line throughout the observation if it occurs during a storm event in one of the counties in the set of counties. Furthermore, the total time during the observation during which there has been a storm event in each county may be determined. The total time during which a line is subject to a storm event is computed by averaging the time during which each county that the line is in has experienced a storm. After that, the storm outage rate is calculated by dividing the number of outages during a storm by the total time it has a storm. Finally, the average storm outage rate is determined by averaging the storm outage rates over all the lines. The non-storm line outage rate and the average non-storm line outage rate are computed similarly.

Based on this analysis, the average non-storm line outage rate is 1.1 and the average storm line outage rate is 8.1. The significant increase in outages during bad weather impacts how historical data is processed and how cascading simulations are performed.

5.6 Visual tracking of the outage and restore process by county

Using the NOAA data and outage data we can track line outages as they move across the system in time and space and project the propagation on the map by county. In Figure 5.4, an identified winter storm from the NOAA storm event dataset is visualized with in counties on the map. The counties are outlined on the map and only counties found in the data set are drawn on the map. The county color is toggled between green (no storm present) and red (storm present) to indicate if a storm is present in that county at that time. as a storm event moves across a region we can track as it moves across BPA in time and in space. The accumulation of line outages is accounted for within each county indicated by a number.

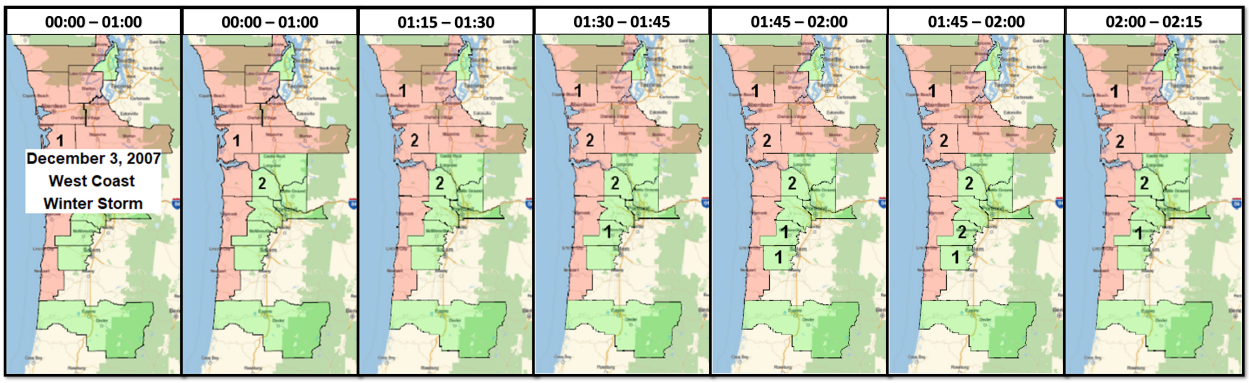

Figure 5.4: County-level map of the West Coast, colored by counties with storms at that hour, and green otherwise. Counties' total outages are shown in still shots over time.

The overall idea of this work is to correlate visually how weather and cascading interact. The interactions show that although a storm may not be present in a county, if a storm and outage is present in a neighboring county, that county may still experience outages before protection kicks in or repairs are made. A video of the still shots for the cascading during the weather can be easily made to watch the progression in time.

5.7 Conclusions

Analyzing the effects of weather on cascading, we analyze historical outage data to identify the origin and propagation of initial outages. Despite analyzing only one large North American utility's 14-year data set, we can draw specific conclusions from the data. Similar data is routinely collected by many utilities worldwide, so the methods can be applied broadly, given access to the data. With a simple process based on outage timing, we can distinguish the initial outage from the subsequent cascade. Weather-related dispatcher cause codes and NOAA storm data are used to study the effects of weather on historical cascading outage data. Although only a few cascades are weather-related, the processing methods used in the chapter show significantly greater propagation from the initial failures and a significantly higher outage rate. In accordance with traditional power system risk analysis, cascading models and analyses will need to recognize and define bad weather and good weather regimes in some way. The traditional power system risk analysis confirms an increased outage rate during bad weather, but the interaction with cascading propagation remains unclear. As a result of processing limitations, the increased outage rate cannot be entirely attributed to cascading effects propagating via network effects. New bulk cascading models and data-processing methods are needed for bad weather conditions. In peak hours and peak months of operation, cascading propagation is less influenced by bad weather, but there is a higher rate of cascades during these peak conditions. Historical outage data is crucial to validating simulations of cascading outages. This chapter presents our bulk statistical data processing methods for historical outage data and NOAA data, which are initial approaches that will be improved in the future. A simple visual of cascading and weather is provided to illustrate the interaction. Nevertheless, our analysis demonstrates the value of this approach for understanding and quantifying key factors in initial outages and cascading, and the prospects for improving methods and gaining further insights are excellent.

5.8 References

[1] I. Dobson, N. Carrington, K. Zhou, Z. Wang, B. Carreras, and J. M. Reynolds Barredo, "Exploring cascading outages and weather via processing historic data," in *Hawaii International Conference on System Sciences 2018*, Big Island, Hawaii, USA, January 2018.

[2] IEEE PES CAMS Working Group on Cascading Failure, "Initial review of methods for cascading failure analysis in electric power transmission systems ieee pes cams task force on understanding, prediction, mitigation and restoration of cascading failures," in *2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century*, 2008, pp. 1–8.

[3] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman, "Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 17, no. 2, p. 026103, 2007.

[4] P. Hines, J. Apt, and S. Talukdar, "Large blackouts in north america: Historical trends and policy implications," *Energy Policy*, vol. 37, no. 12, pp. 5249–5259, 2009.

[5] D. E. Newman, B. A. Carreras, V. E. Lynch, and I. Dobson, "Exploring complex systems aspects of blackout risk and mitigation," *IEEE Transactions on Reliability*, vol. 60, no. 1, pp. 134–143, 2011.

[6] B. A. Carreras, D. E. Newman, and I. Dobson, "North american blackout time series statistics and implications for blackout risk," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4406–4414, 2016.

[7] IEEE PES CAMS Working Group on Cascading Failure, "Initial review of methods for cascading failure analysis in electric power transmission systems ieee pes cams task force on understanding, prediction, mitigation and restoration of cascading failures," in *2011 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century*, 2011, pp. 1–8.

[8] IEEE PES CAMS Working Group on Cascading Failure, "Benchmarking and validation of cascading failure analysis tools," *IEEE Trans. Power Systems*, vol. 31, no. 6, pp. 4887–4900, November 2016.

[9] B. A. Carreras, D. E. Newman, I. Dobson, and N. S. Degala, "Validating OPA with wecc data," in *2013 46th Hawaii International Conference on System Sciences*. Maui, HI USA,: IEEE, January 2013, pp. 2197–2204.

[10] M. Papic and I. Dobson, "Comparing a transmission planning study of cascading with historical line outage data," in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, 2016, pp. 1–7.

[11] M. A. Rios, D. S. Kirschen, D. Jayaweera, D. P. Nedic, and R. N. Allan, "Value of security: modeling time-dependent phenomena and weather conditions," *IEEE Transactions on Power Systems*, vol. 17, no. 3, pp. 543–548, 2002.

[12] E. Ciapessoni, D. Cirio, G. Kjølle, S. Massucco, A. Pitto, and M. Sforna, "Probabilistic risk-based security assessment of power systems considering incumbent threats and uncertainties," *IEEE Transactions on Smart Grid*, vol. 7, no. 6, pp. 2890–2903, 2016.

[13] F. Cadini, G. L. Agliardi, and E. Zio, "A modeling and simulation framework for the reliability/availability assessment of a power transmission grid subject to cascading failures under extreme weather conditions," *Applied energy*, vol. 185, pp. 267–279, 2017.

[14] R. Yao and K. Sun, "Towards simulation and risk assessment of weather-related cascading outages," *arXiv preprint arXiv:1705.01671*, 2017.

[15] "Bonneville Power Administration transmission services operations & reliability website." [Online]. Available: https://transmission.bpa.gov/Business/Operations/Outages

[16] North American Electric Reliability Corporation, "Transmission availability data system (tads) data reporting instruction manual," North American Electric Reliability Corporation, Tech. Rep., August 2014.

[17] J. J. Bian, S. Ekisheva, and A. Slone, "Top risks to transmission outages," in *2014 IEEE PES General Meeting*, 2014, pp. 1–5.

[18] National Oceanic and Atmospheric Administration (NOAA), "National centers for environmental information storm events database." [Online]. Available: https://www.ncdc.noaa.gov/stormevents

[19] I. Dobson, B. A. Carreras, D. E. Newman, and J. M. Reynolds-Barredo, "Obtaining statistics of cascading line outages spreading in an electric transmission network from standard utility data," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4831–4841, November 2016.

[20] I. Dobson and D. E. Newman, "Cascading blackout overall structure and some implications for sampling and mitigation," *International Journal of Electrical Power & Energy Systems*, vol. 86, pp. 29–32, 2017.

[21] H. Ren and I. Dobson, "Using transmission line outage data to estimate cascading failure propagation in an electric power system," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 55, no. 9, pp. 927–931, 2008.

[22] I. Dobson, "Estimating the propagation and extent of cascading line outages from utility data with a branching process," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 2146–2155, November 2012.

[23] R. Billinton and G. Singh, "Application of adverse and extreme adverse weather: modelling in transmission and distribution system reliability evaluation," *IEE Proceedings-Generation, Transmission and Distribution*, vol. 153, no. 1, pp. 115–120, 2006.

[24] R. Billington and R. N. Allan, *Reliability evaluation of power systems.* Plenum Publishing Corp., New York, NY, 1984.

CHAPTER 6. DATA ANALYSIS TOOL FOR CONSUMPTION DATA FROM A DISTRIBUTION UTILITY

John Bilsten, Algona Municipal Utilities, Algona, Iowa, USA

Anne Kimber, Electric Power Research Center, Iowa State University, Ames, Iowa, USA

Nichelle'Le K. Carrington and Zhaoyu Wang, Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa, USA

Modified from the user guide for Customer Clustering using AMI tool for Small Public Utilities in American Public Power Association's DEED Project Library [1].

6.1 Abstract

We have created software for small utilities that processes and analyzes data from advance metering infrastructure (AMI). AMI data is a recording of energy use for distribution customers at the building level that can be recorded down to the minute. Small utilities can use a deployable tool to help them handle AMI data. The capacity of categorizing clients based on consumption is one of the tool's analyses. Customers are grouped based on load in this analysis using k-means clustering. Another analytical option in the program is a detailed breakdown of the load usage. Each client class's contribution to the hourly load usage is broken below. The program was created to allow tiny utilities to clean, analyze, and export data.

6.2 Overview

Utilities, including public power (defined as consumer owned and governed by city councils or municipal utility boards) have upgraded their data collection methods and processing capability by deploying new technology such as smart meters and advanced metering infrastructure (AMI) [2–4] These investments may be made to enable utilities to better detect and respond to

distribution outages. AMI data can also enable utilities to better understand and respond to customer consumption patterns, and especially contribution to peak consumption [5, 6]. Stochastic and probabilistic data-driven modeling techniques can be used to extract consumption patterns from historical data to make strategic decisions from the readings. Utilities using data-driven modeling techniques find helpful information that can characterize their customers based on load consumption and validate the existing rate class classifications (commercial, residential, etc.). This understanding can allow the utility to plan investments in new rate designs, demand response programs for peak shaving, energy efficiency assistance, renewable energy project locations, and distribution system infrastructure planning.

Small utilities that invest in AMI may have few staff members who have the time and skills needed to extract valuable information from AMI data. The extremely large data files are an obstacle to analysis. The current practice of utilities is to review the monthly energy consumption based on billing for rate class which makes it challenging to define the daily behavior of customers [3]. Exploring historical consumption data to gain insight into customers' load behavior has significant benefits to small utilities with a small staff.

This chapter presents the components used to develop a research-grade, Excel-based software tool that small public utilities can use to extract useful information from AMI-based data. The purpose of the software tool is to help small utilities process AMI-based data with limited staff resources and save on data processing expenses. The tool can import, process, analyze, and export large volumes of AMI data recorded at intervals of 15-minutes and 60-minutes. Currently, there are consulting services offered commercially, but there is no research-grade tool available to complete such tasks for small utilities. The organization of this chapter is as follows. Section 6.3 presents a literature survey of current techniques for analyzing smart meter data within the field. Section 6.4 describes the consumption data provided by a small distribution utility located in the US. Section 6.5 defines and highlights the data mining techniques in the data management plan used for processing, saving, and exporting data using the tool. Section 6.6 gives the details of the mathematical formulas used for the five analyses within the tool. Section 6.7 presents the

architecture of how the tool was structured using the details from the previous sections. Finally, Section 6.8 presents the conclusion that summarizes the entire chapter.

6.3 Literature Survey

Advanced Metering Infrastructure (AMI) is a critical network that consists of smart meters, communication, and network support for two-way communication between the utility and the consumer [6]. An AMI-based system's consumption data can consist of multiple measurements at varying time granularity, making the dataset very heterogeneous, voluminous, and challenging to manage. How to handle the data and determine its integrity are a couple of the many issues that arise when processing consumption data to extract information [5]. A clean uniform data set must be produced to extract information from historical consumption data.

In many works in energy literature, the use of data partitioning methods increase the manageability of data for processing and performing analysis [7–13]. The primary focus of these works is the comparison of different methods for producing load profiles and validating how accurately the profiles capture the variations in electricity demand across customers. They do not partition data by customer class. In [14], a segmentation of AMI data was achieved using AMI data but was grouped by season and not by customer class. A common practice to understand customers' load behavior is to apply demand profiles derived from clustered data. Balachandra and Chandru identified nine representative load curves obtained through clustering load profiles in a system planning exercise for Karnataka in India[10]. Chicco presented a detailed comparison of the performance of several clustering algorithms that results showed that the modified follow-the-leader and the hierarchical clustering was the most effective and suitable algorithms for customer clustering [15]. Marton et al. demponstrated order-specific clustering algorithm on a case study using electricity demand data[16]. However, these works did not attempt to identify relevant characteristics of consumption from within the data as a start for clustering consumers. Instead, each work used a predefined set of characteristics, not customer data.

Several works note that utilities rely on responses from customer surveys to gain insight into consumer behavior. Customer surveys are similar to the psycho-graphic consumer segmentation based on customers' feelings and actions presented in [17]. However, their segmentation is based solely on survey data about consumers' behavior and attitude toward electricity and energy conservation and did not involve processing any consumption data. The study in [18] proposed procedures to develop probabilistic load models observed from a distribution-level feeder of a residential community. The paper does not use AMI data and only models the uncertainty of aggregated feeder-level loads. A Matlab-Simulink-GUIDE tool focusing on the simulation of residential load at the appliance level was proposed in [19]. Still, it did not consider the load of multiple houses or different customer classes. References [2–4, 20, 21] explore different techniques of clustering AMI data for distribution systems specifically residential for stratification. The benefit of leveraging AMI data for analysis, demand forecasting, and state estimation is presented in [22–24].

The above literature has certain limitations. Firstly, the existing papers use simulated data and publicly available AMI data of specific customers. Therefore, their data may not be sufficient to capture load behaviors of different customer classes and is not representative of small municipal utilities. Furthermore, the existing research focuses more on analyzing residential customers. The load behaviors of other classes of customers have not been thoroughly studied. None of the listed resources developed a deployable, standalone tool for their model. In addition, to the best of our knowledge, no similar software tool has been reported in the literature. For this tool, an actual U.S. small municipal utility provided our dataset with different customer categories.

6.4 AMI Data Description

The AMI data used in this chapter are from a small public power utility that serves roughly 5,000 customers and were used to develop the structure of data in the standalone tool in section 6.7. The utility divides its customers into seven rate classes: residential, small commercial, large commercial, and industrial, public authorities, public street lighting, and school lighting and fair-

grounds. Customers are served on several feeders. All rate class have a customer charge and energy charge based on their expected usage, with a demand charge for large commercial 15-minute demand (greater than 20 kW) and industrial customers (demand greater than 250 kW). Public authorities (e.g. power supplied to municipally-owned facilities) are charged at small commercial, large commercial, or industrial rates depending on demand.

Within the public utility system, over 3,000 customers have smart meters installed, including industrial customers. The utility and its subcontractors provided three data file types. The meter data management contractor provided monthly comma-separated values (.CSV) files of consumption data with the multipliers applied. The consumption data was recorded in kWh and the accompanying account information of customer meter and account information were provided for cross-reference to meter consumption data. Data was recorded at either hourly or 15 minute intervals (for industrial customers). The utility collected consumption data for four years from 2014 to 2018.

6.5 Data Mining

Data mining is defined as the process that integrates a data management plan to extract, process, and obtain helpful information from a given data set [3, 25]. An effective data management plan aims to process, store, and clean data properly. Understandable models use clean data to allow the discovery of patterns and trends and the behavior of loads. [3]. In this work, the AMI tool incorporated a data management plan with six steps[3, 25]:

a) **Storing Data-** the raw, processed, and supporting information files are stored in a dedicated file folder location.

b) **Preprocessing Data-** uses data mining techniques that aid in converting raw data files into an understandable processed global table for analysis in the tool.

c) **Analyzing Data-** involves systematically applying statistical and logical strategies to describe, summarize, illustrate and evaluate these summary data organized by the rate code (customer class), feeder, or time options (hour, day, week).

d) **Preserving Data-** data is preserved within the tool by storing summary data in a global table and calling from the global table to perform each analysis independently from one another.

e) **Export Data-** processed data is exportable from the tool as selected tables and pictures of graphs. The tables are saved as Excel files and imported into Word, NotePad, WordPad, or Excel. The pictures are .png files that can be imported into Word. The pictures of the graphs can be used for reports or descriptive data for the loads.

f) **Reuse Data-** a unique feature is that the processed exported data files from the previous analysis can be re-imported into the tool for further analysis. This is a convenient feature that allows for faster processing within the tool.

As seen in Figure 6.1, the diagram of the data management plan illustrates how the data flow throughout the tool.

Figure 6.1: Diagram of the data management plan

The benefit of this data management plan is that it allows for voluminous, raw data to be cleaned to remove incomplete data entries. After cleaning, simple statistical analyses are performed on the data set and expressed in simple summary tables and graphs. A unique feature of this tool is that the modified data management plan inclusion allows the processed data to be reused in the tool and outside the tool. Any data that has been cleaned using the tool or generated from an analysis in the tool can be saved as an comma-separated values (.CSV), text (.TXT), or Microsoft Excel spreadsheet file (.XLSX).

6.6 AMI Data Analysis

Load duration curve, load profile, customer contribution, customer statistics, and customer classification are the five analyses provided by the tool. This section defines the formulas for these analyses. The analyses in the tool are equipped to process and display results for an hour, day, week, month, or season. A case study for a summer day, July 17, 2017, was included in this section to demonstrate the output from the analyses.

6.6.1 Load Profile

A load profile is the energy consumption pattern for a customer or group of customers over a given period, such as hour, day, week, month, season, and year [26]. It is presented as a two-dimensional graphical display of the aggregated load of a specific group of customers over a given period at an hourly interval. The aggregated load $L(t)$ at time t is defined by the summation of the total load of each customer for that rate class as

$$L(t) = \sum_{i=1}^{C} L_i \tag{6.1}$$

C is the total number of customers for that customer rate class. $L(t)$ is calculated for each hour for each rate class for 24 hours. In the tool the load profile for each rate class is constructed by plotting the $L(t)$ versus the time in chronological order shown in Figure 6.2.

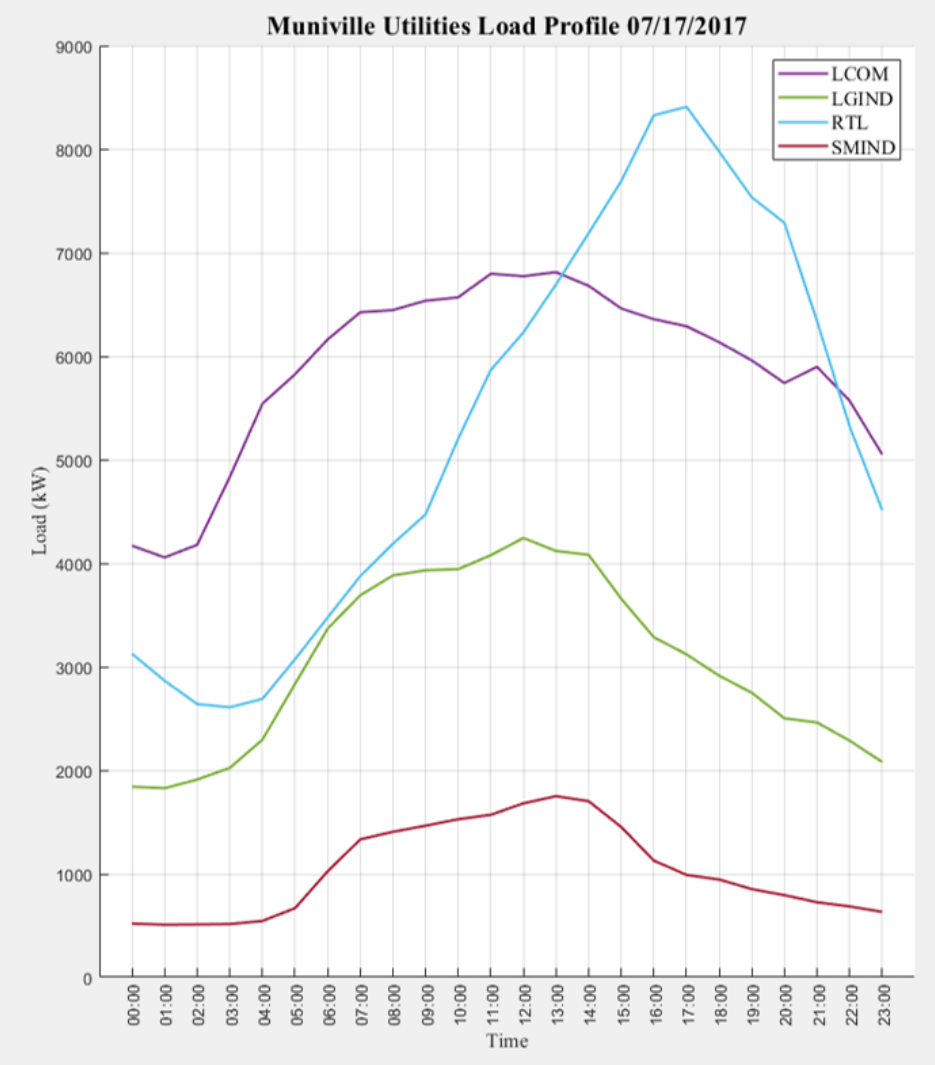

Figure 6.2: Each line in the plot represents the aggregated load for the day for that class.

6.6.2 Load Duration

A load duration curve illustrates the aggregated load for each hour used in the load duration analysis, and is calculated using 6.1.

Unlike a load profile, the load in the load duration curve is plotted in descending order and not chronological order. The average load is calculated by

$$\overline{L} = \frac{1}{C} \sum_{i=1}^{C} L_i \tag{6.2}$$

Let $\overline{L}$ be the average load for the entire observed period, and C is the total number of customers. The tool plots the load duration curve as a graphical display of the sorted aggregated load against the average load $\overline{L}$. The average load $\overline{L}$ is also shown in Figure 6.3.

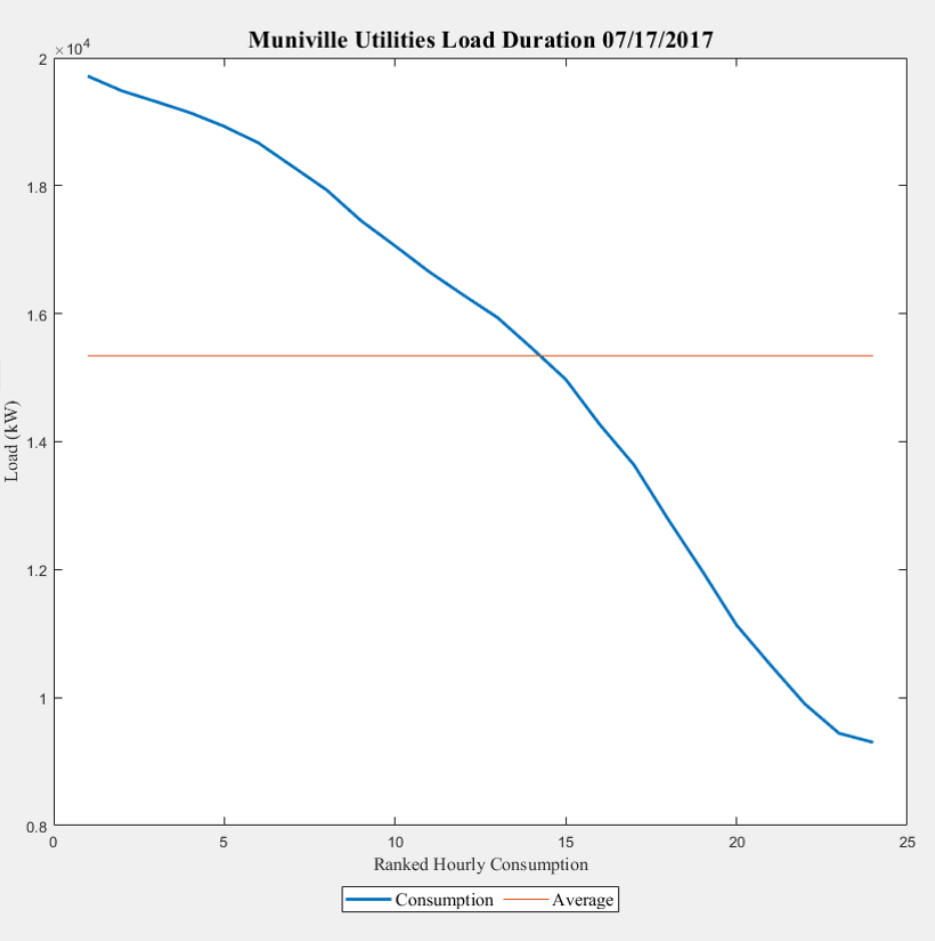

Figure 6.3: Load duration curve for day using all classes.

6.6.3 Customer Contribution

The Customer Contribution is the sum of the loads for all customers within a rate code for the time period selected for the rate code, either an hour or a 15-minute interval.

$$\sum_{i=1}^{R} L(t)_i \tag{6.3}$$

The stacked bar graph displays the components of total load for each hour of a selected day or a week. In Figure 6.4, the aggregated load for each rate class for a day is displayed as a stacked bar graph.

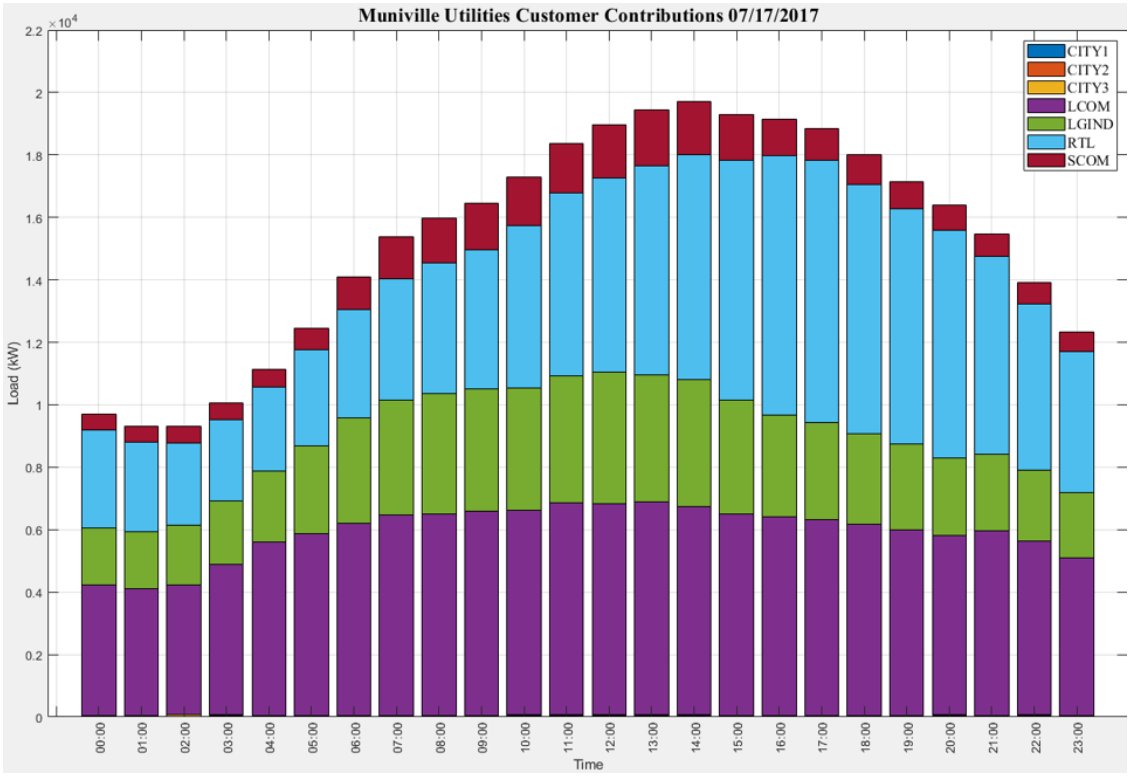

Figure 6.4: Each colored segment is the total consumption of that rate code

Each segment of a stacked bar represents the total consumption for that rate class for the hour.

6.6.4 Customer Statistics

Basic statistics such as minimum (min), maximum (max), median, and standard deviation (SD) are commonly used to describe data sets. Max and min values indicate the largest and

smallest value in the dataset of a rate class, respectively. The mean of a rate class is calculated using (6.2) and the standard deviation (SD) is calculated using

$$S = \sqrt{\frac{1}{C-1}\sum_{i=1}^{C}|L_i - \overline{L}|^2} \tag{6.4}$$

Where S is the square root of the variance of the mean $\overline{L}$ of that rate class. The customer statistics in the tool is illustrated using box and whisker plots. In Figure 6.5, the consumption for each customer class has its own box. The max and min load are the upper and lower loads of the customer class represented with a whisker. The SD is illustrated with the length of the whisker, if the whisker is long the SD is high and low if the whisker is small. The median load is represented with the red line in the box.

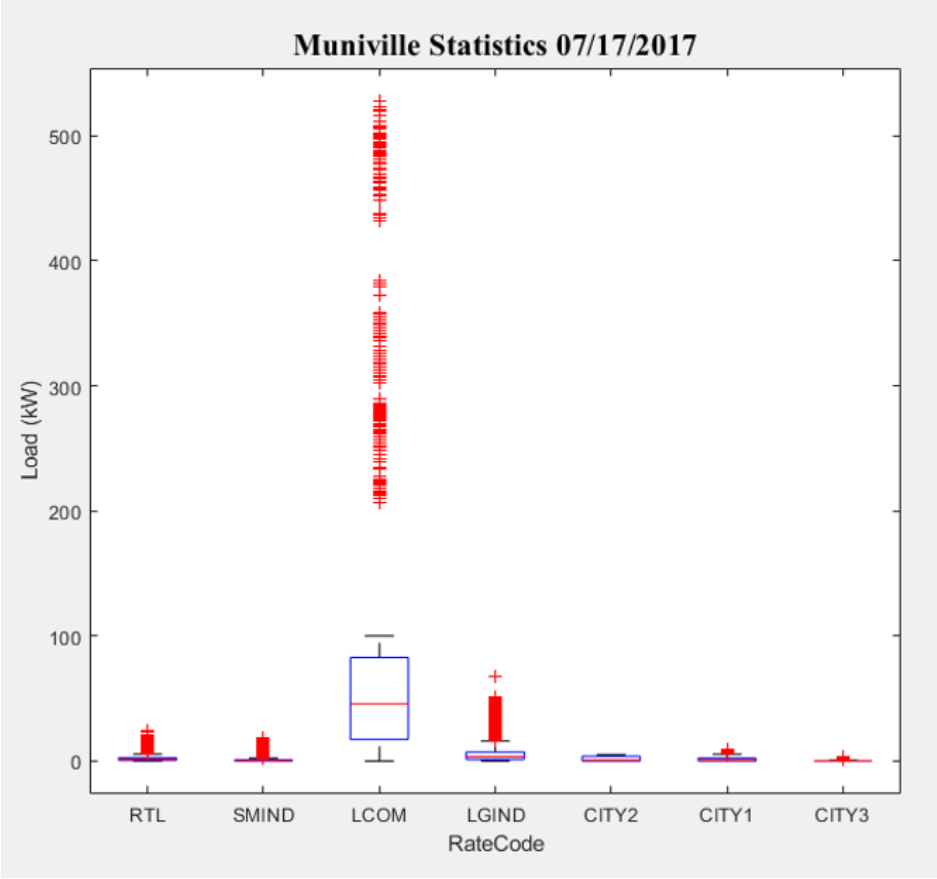

Figure 6.5: Box plot for each customer class.

The red points in Figure 6.5 the plots are the outliers for that class. In Figure 6.5 the large range in comsumption for the Large Commercial rate class shows that these customers are very diverse and may need more exploration of the customer consumption behavior.

6.6.5 Clustering Analysis

Although a customer is associated with a specific rate class, their consumption may not reflect the typical behavior of the group. Classifying customers based on consumption helps utilities identify outliers for evaluating rate class placement and design of demand charges. The MatLab ®environment uses the k-Means++ algorithm for clustering analysis [27]. The k-Means++ algorithm was developed by [28] which is a modified version of the k-Means algorithm also known as Lloyd's algorithm [29] to achieve a faster convergence to a sum of within-cluster than its predecessor. It uses a two-phase, heuristic process to find k centroids, partition the data into one of the k clusters using said centroids. In the tool, the algorithm creates k centroids from the average load $\overline{L}$ of each customer class for the period specified (hour, day, week, or month) to initialize. Let X be the data set of customers, and each centroid be c_i for $i \geq k$. The initial centroid c_1 selection for a data point x_i is assigned at random, and the standard distance between the two is computed as $d(x_i, c_1)$. The distance of the data points to a second centroid c_j is then computed. The selection of a second centroid for a random load in X is made by computing the probability using

$$\frac{d^2(x_i, c_1)}{\sum_{j=1}^{n} d^2(x_j, c_1)} \tag{6.5}$$

where n is the total number of observations in X. In the second phase, the algorithm then computes the distance of each observation x_j to each centroid and selects the closest as the new centroid. The next step is a repetitive calculation to select centroid j at random of the probability

$$\frac{d^2(x_m, c_p)}{\sum_{h;x_h \in C_p} d^2(x_h, c_p)} \tag{6.6}$$

for $m = 1, \ldots, n$ and $p = 1, \ldots, j–1n$. Let C_p be the set of all observations closest to centroid c_p and x_m belongs to C_p. This step is repeated until k centroids are chosen. The updated centroid for each category are plotted with the mean cluster load for the rate class to compare clustering results as shown in Figure 6.6.

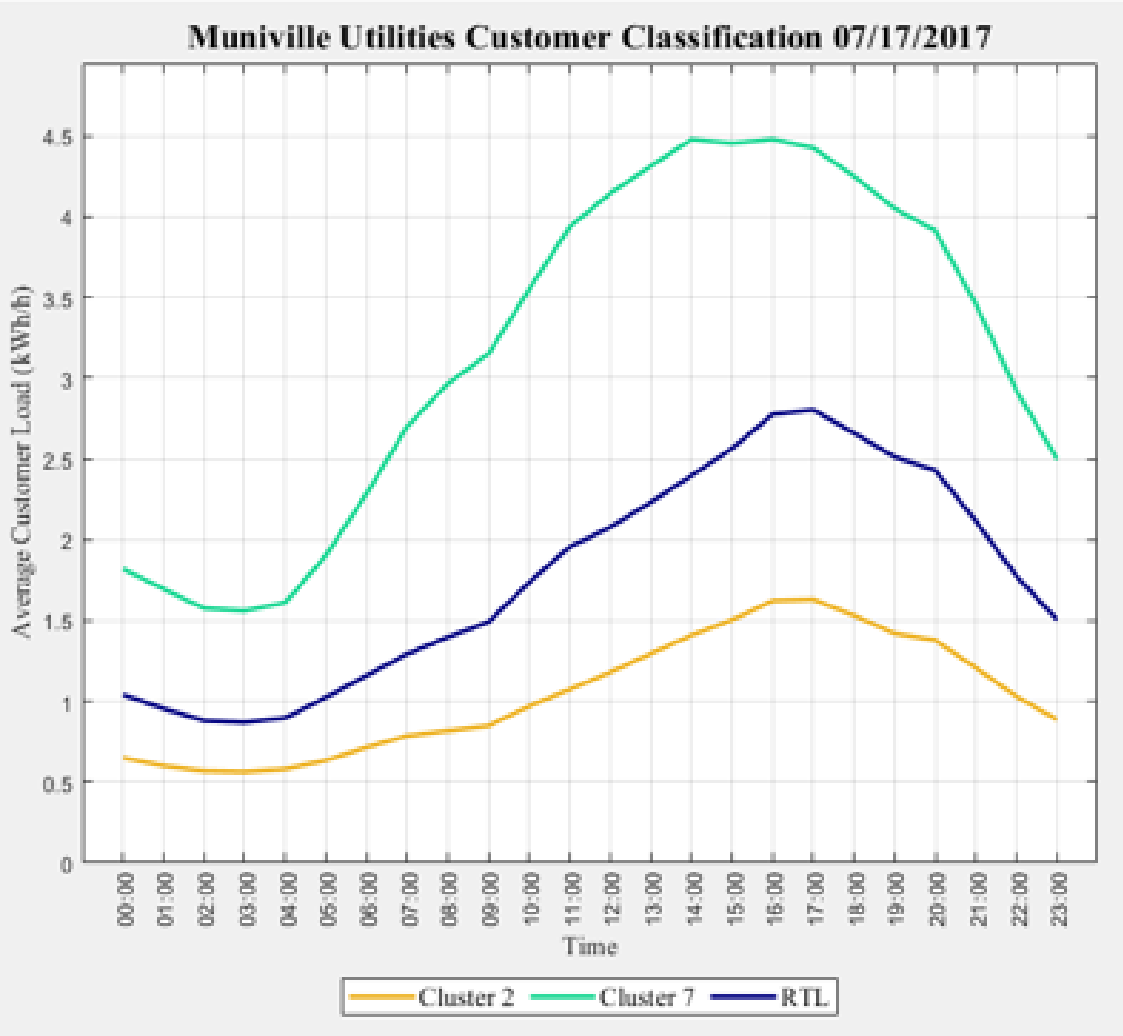

Figure 6.6: Cluster loads that contained the majority of residential customers plotted with the average of the residential load.

The sample graph in Figure 6.6, shows the average individual load for the selected customer class (purple), the average individual load of the dominant cluster (green), and the average in-

dividual load of all the other clusters containing at least 15% of the customers from the selected rate class (yellow). A dominant cluster for a given rate class is a cluster containing the majority of the customers in that rate class.

6.7 AMI Data Mining Tool

The architecture of the tool has four stages that incorporate the data management plan for data mining. The four stages are Data Import, Data Preprocessing, Data Analysis, and Data Export. These four stages follow the data management plan to process and deliver uniform data files to the user. The flow chart of the tool's architecture is a visual representation of how the data is formatted and accessed by the tool within the individual processes as a flowchart. This architecture will be presented in the stage descriptions below.

6.7.1 Stage 1: Data Import

The architecture of the Data Import stage in the flowchart of Figure 6.7 allows two methods of importation of AMI data within the tool. The first option is to merge individual "raw" text and Excel files into one table. Raw data refers to individual files of AMI consumption, feeder, account information, and heating/cooling/temperature data that have not been merged. The second option is to import "processed data" that has been cleaned and formatted by the tool from a previous analysis. Processed data refers to a combined file of individual AMI consumption, feeder, account information, and heating/cooling/temperature data.

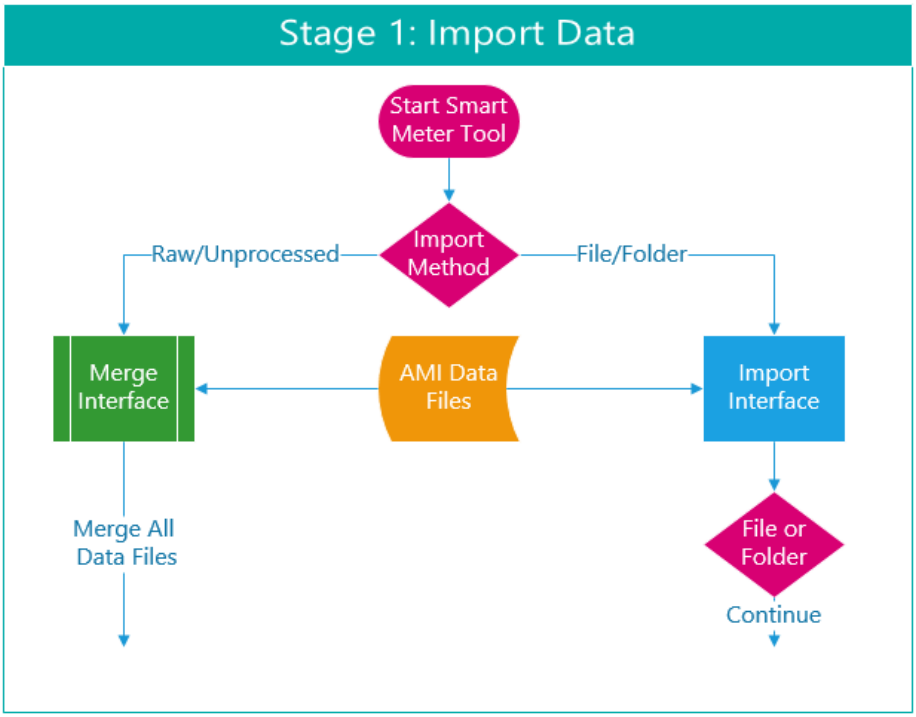

Figure 6.7: Flowchart of options for importing data into the tool.

In Figure 6.7, data files are imported as either raw or previously processed data files exported from the tool. The raw data is processed using subroutines to format the data into a combined uniform information table which is then saved in a datastore internally within the tool. The combined data from the Stage 1 will be stored in a datastore within the tool as a tall array.

6.7.2 Stage 2: Data Preprocessing

In Stage 2, the data is imported and cleaned. This is the main purpose of the tool. The flow chart for Stage 2 is shown in Figure 6.8. This shows how the data steps through the cleaning process. This step helps create the data into a global table that can be called from any analysis process in the tool. The global table allows the results of each analysis to be independent from the datastore, saved independently, and available for further use within and outside of the tool.

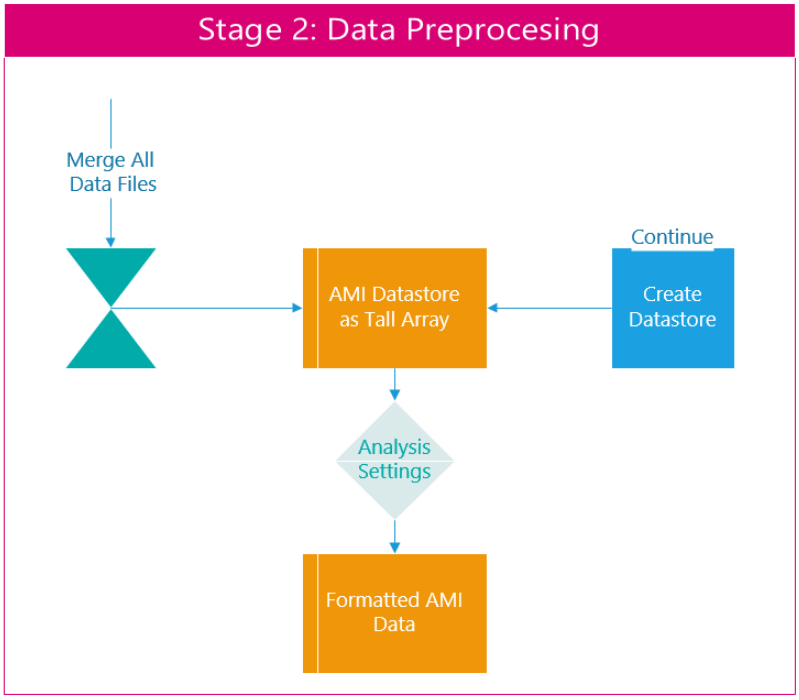

Figure 6.8: Flowchart of process for formatting and filtering data for analysis within the tool.

The global table is created by filtering the data using the analysis settings. Filtering by combinations of date, feeder, customer class, hourly intervals, or weekly information aid in mining data. The condensed data is saved and stored for reuse.

6.7.3 Stage 3: Data Analysis

In Stage 3, the processed data from Stage 2 can be evaluated with five independent analyses: customer classification, customer contribution, load profile, load duration, and customer statistics. The customer classification screen has a subroutine that groups customers by rate code using the unsupervised learning technique K-Means to group the customers in categories of similar consumption for the period as in [3]. The Analysis menu is used to access the analysis procedures and other interfaces except Stage 1. from the analysis menu on the Analysis Menu. The main screen serves as a centralized hub to access and review the results from each analysis process.

Each analysis subroutine calls upon the datastore generated from Stage 2 to evaluate the supporting algorithm for a subroutine.

The Customer Contribution screen has a subroutine that displays and calculates a rate code's total contributions within a given period at either the hour or 15-minute interval. The Load Profile screen has a subroutine that graphs the variation in the electrical load versus time. The Load Duration screen has a subroutine that sorts and graphs the load in descending order of magnitude as a load duration curve.

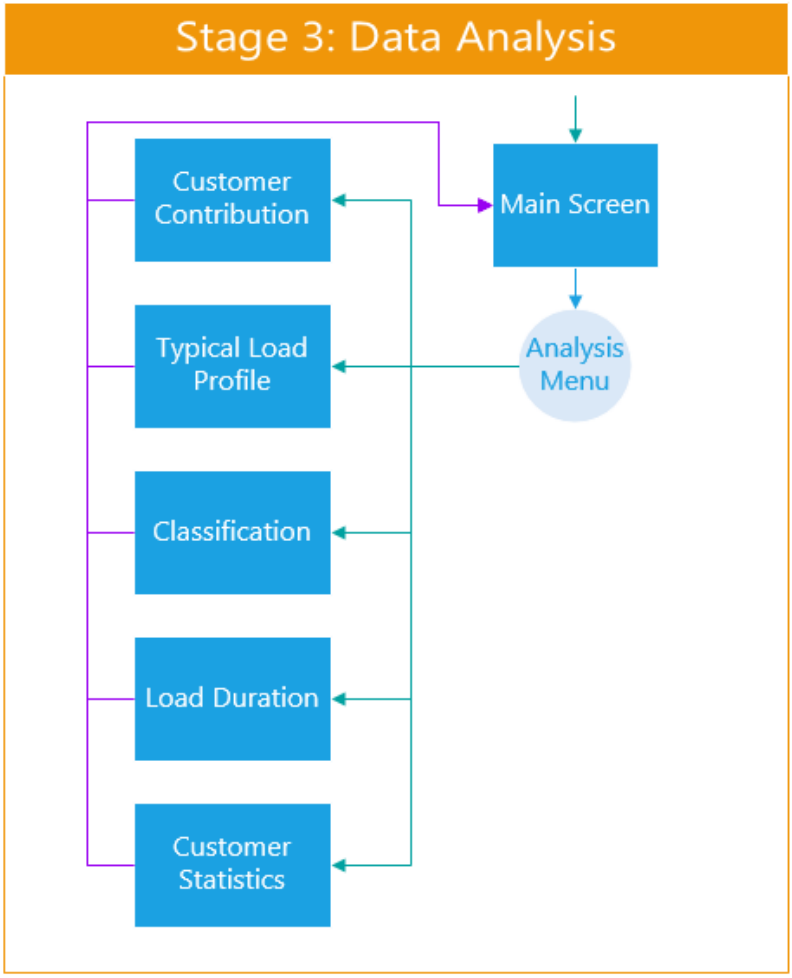

Figure 6.9: Flowchart of options for analyzing the data within the tool.

6.7.4 Stage 4: Data Export

In Stage 4: Data Export, the user can export processed data and the graphs from analyses as tables and figures through the Export menu as shown in Figure 6.10.

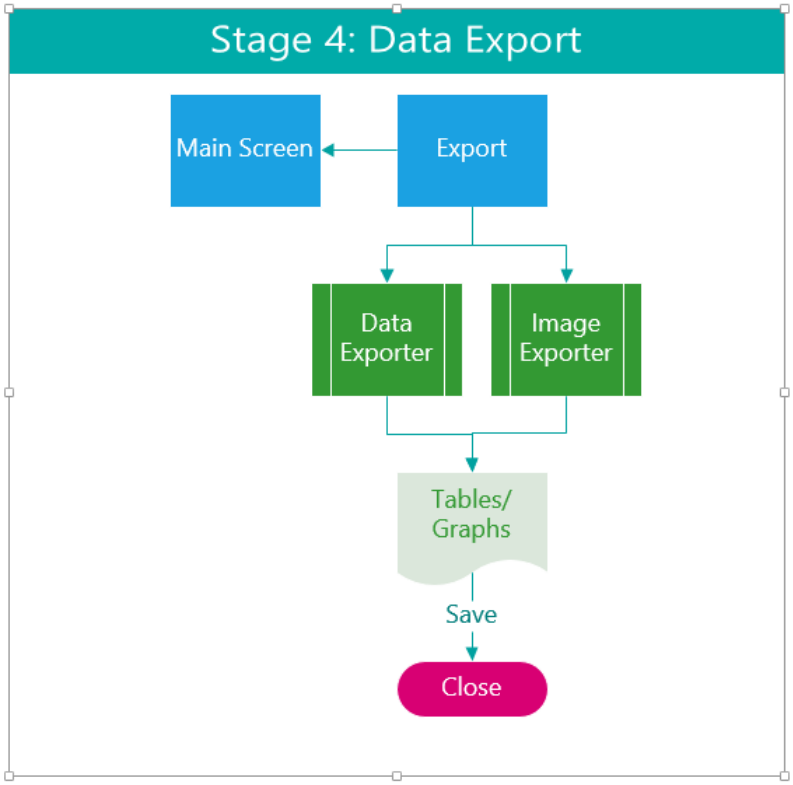

Figure 6.10: Flowchart of the within the tool options for importing data.

The Data Exporter screen allows users to save the tables from the tool as a .csv, .xlsx, or .txt extension file to their computer. The Data Exporter screen enables the user to save the figures from each analysis as a Portable Network Graphic (.png) file.

6.8 Conclusion

The work described in this chapter produced a research-grade tool for small utilities with features that import, process, analyze and export tables and graphs of AMI data. The tool gives small utilities the ability to study customer behavior by feeder, rate class and individual cus-

tomers for a day or week. Load profiles and load duration curves offer visuals of the aggregated load. Clustering analysis segments the customers into groups based on their consumption for a day, a week.

During the development of this tool, we and the small public utility did extensive review and error-checking of AMI data sets, comparing these with supervisory control and data acquisition (SCADA) data summaries and the utility's monthly reports. Some data sets had long sequences of repeated data that did not match actual customer behavior. This issue was a problem in the early years when AMI meters had been deployed. The use of a simple tool, like the one developed in this project, would enable utilities to scan data and find odd outliers that could impact billing data or incorrectly indicate outages that did not actually occur because of artifacts from AMI meter communication errors.

Utilities planning to deploy AMI meters will need a plan executed with billing and meter data management contractors to retrieve and analyze the data. This plan would ensure that the data will be available, formatted, and have meter multipliers applied.

By exploring the AMI data, utilities can gain insights into their customer load behavior and the main determinants affecting load consumption. Customer groupings based on load characteristics and time-varying probabilistic distributions of load consumption can enable various higher-level applications such as usage-specific tariff structures, consumer-specific demand response programs, cost/benefit analysis of renewable energy integration programs, and conservation voltage reduction.

6.9 References

[1] *Customer Clustering using AMI tool for Small Public Utilities.* [Online]. Available: https://www.publicpower.org/deed-project/customer-clustering-using-ami-tool-small-public-utilities

[2] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, "Customer classification and load profiling method for distribution systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1755–1763, 2011.

[3] J. E. Parra, F. L. Quilumba, and H. N. Arcos, "Customers' demand clustering analysis—a case study using smart meter data," in *2016 IEEE PES Transmission & Distribution Conference and Exposition-Latin America (PES T&D-LA)*. IEEE, 2016, pp. 1–7.

[4] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE transactions on smart grid*, vol. 7, no. 1, pp. 136–144, 2016.

[5] A. Ghosal and M. Conti, "Key management systems for smart grid advanced metering infrastructure: A survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2831–2848, 2019.

[6] C. Selvam, K. Srinivas, G. Ayyappan, and M. Venkatachala Sarma, "Advanced metering infrastructure for smart grid applications," in *2012 International Conference on Recent Trends in Information Technology*, 2012, pp. 145–150.

[7] S. Lin, F. Li, E. Tian, Y. Fu, and D. Li, "Clustering load profiles for demand response applications," *IEEE Transactions on Smart Grid*, 2017.

[8] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 420–430, 2014.

[9] Y. R. Gahrooei, A. Khodabakhshian, and R.-A. Hooshmand, "A new pseudo load profile determination approach in low voltage distribution networks," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 463–472, 2018.

[10] P. Balachandra and V. Chandru, "Modelling electricity demand with representative load curves," *Energy*, vol. 24, no. 3, pp. 219–230, 1999.

[11] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Consumers' load profile determination based on different classification methods," in *2003 IEEE Power Engineering Society General Meeting*, vol. 2. IEEE, 2003, pp. 990–995.

[12] M. Beccali, M. Cellura, V. L. Brano, and A. Marvuglia, "Forecasting daily urban electric load profiles using artificial neural networks," *Energy conversion and management*, vol. 45, no. 18-19, pp. 2879–2900, 2004.

[13] M. Hoffmann, L. Kotzur, D. Stolten, and M. Robinius, "A review on time series aggregation methods for energy system models," *Energies*, vol. 13, no. 3, p. 641, 2020.

[14] Y. Yuan, K. Dehghanpour, F. Bu, and Z. Wang, "A data-driven customer segmentation strategy based on contribution to system peak demand," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4026–4035, 2020.

[15] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions on Power Systems*, vol. 21, no. 2, pp. 933–940, 2006.

[16] C. Marton, A. Elkamel, and T. A. Duever, "An order-specific clustering algorithm for the determination of representative demand curves," *Computers & Chemical Engineering*, vol. 32, no. 6, pp. 1365–1372, 2008.

[17] P. Balachandra and V. Chandru, "Supply demand matching in resource constrained electricity systems," *Energy Conversion and Management*, vol. 44, no. 3, pp. 411–437, 2003.

[18] A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4019–4030, 2013.

[19] E. Pouresmaeil, J. M. Gonzalez, C. Canizares, and K. Bhattacharya, "Development of a smart residential load simulator for energy management in smart grids," *IEEE Transactions on Power Systems*, pp. 1–8, 2013.

[20] R. Green, I. Staffell, and N. Vasilakos, "Divide and conquer? k-means clustering of demand data allows rapid and accurate simulations of the british electricity system," *IEEE Transactions on Engineering Management*, vol. 61, no. 2, pp. 251–260, 2014.

[21] R. Li, F. Li, and N. D. Smith, "Multi-resolution load profile clustering for smart metering data," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4473–4482, 2016.

[22] J. Peppanen, M. J. Reno, M. Thakkar, S. Grijalva, and R. G. Harley, "Leveraging ami data for distribution system model calibration and situational awareness," *IEEE transactions on smart grid*, vol. 6, no. 4, pp. 2050–2059, 2015.

[23] B. Hayes, J. Gruber, and M. Prodanovic, "Short-term load forecasting at the local level using smart meter data," in *2015 IEEE Eindhoven PowerTech*. IEEE, 2015, pp. 1–6.

[24] V. Mtembo, G. A. Taylor, and A. Ekwue, "A novel econometric model for peak demand forecasting," in *2014 49th International Universities Power Engineering Conference (UPEC)*. IEEE, 2014, pp. 1–6.

[25] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[26] D. Gerbec, S. Gasperic, and F. Gubina, "Determination and allocation of typical load profiles to the eligible consumers," in *2003 IEEE Bologna Power Tech Conference Proceedings,*, vol. 1. IEEE, 2003, pp. 5–pp.

[27] I. The MathWorks, *Statistics and Machine Learning Toolbox-Cluster Analysis:kmeans*, Natick, Massachusetts, United State, 2019. [Online]. Available: https://www.mathworks.com/help/stats/kmeans.html?s_tid=doc_ta

[28] S. Vassilvitskii and D. Arthur, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2006, pp. 1027–1035.

[29] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

CHAPTER 7. TRANSMISSION GRID OUTAGE STATISTICS EXTRACTED FROM A WEB PAGE LOGGING OUTAGES IN NORTHEAST AMERICA

Nichelle'Le K. Carrington, Ian Dobson, and Zhaoyu Wang, Department of Electrical and Computer Engineering Iowa State University, Ames, Iowa, USA

Modified from a manuscript published in The 53^{rd} North American Power Symposium (NAPS 2021) [1]

7.1 Abstract

Detailed outage data is foundational for the study of power transmission grid reliability and resilience, and particularly for dependent outages and rarer events, but there are very few such data sets that are published and freely accessible to all engineers and researchers. There are voluminous logs of scheduled and actual outages in a region of Northeast America available on the web. We show how to compress and process these logged data to obtain bulk statistics describing the outages, such as event size, propagation, and spread on the network. These statistics are very useful for calibrating and validating models of resilience to ensure realism, and in developing data-driven approaches.

7.2 Overview

Obtaining detailed outage data is difficult for researchers and engineers due to confidentiality restrictions and the sensitivity of the information once processed. The inclusion of detailed outage data is foundational for engineering models and simulation for reliability and resilience. Outage rates averaged over classes of equipment and periods are published by some national organizations such as [2]. Many books and chapters have realistic annual outage rates for specific

test systems and sometimes for broad categories of weather conditions. All these averaged, and typical data are helpful, especially for steady-state Markov modeling of reliability and detecting trends in reliability. However, it averaged for many problems involving dependencies within outages and rare events, including common cause outages, cascading, and resilience. Typical single component data do not suffice. The timings and details of many specific outages are required to advance the field.

While it is sometimes feasible for engineers and researchers to gain access to detailed industry outage data with non-disclosure agreements and publish some suitably non-identifying overall results, there is a unique role for public data in advancing the field since methods based on public data can be reproduced and improved on by other investigators. Moreover, the developed methods can subsequently be applied across the industry since transmission utilities, and system operators in North America and worldwide routinely collect their detailed outage data.

There are few detailed outage records freely available to engineers and researchers; indeed, to the authors' knowledge, there has been only one such source for transmission line outage data, namely the Bonneville Power Administration (BPA) website [3]. In this chapter, we show how to obtain detailed outage data from a second public website.

The BPA published data has been processed in various ways in [4–6] and used to validate and calibrate models in [5, 7–11]. There are many potential applications for detailed outage data, and we seek to generally facilitate applications by showing how to extract the new data. One application studies the size, propagation, and spread of outages that bunch together in cascading or weather-induced events, and we show the bulk statistics that can be obtained from the detailed outage data. These bulk statistics help calibrate and validate cascading models and simulations [12, 13], or can be sampled to drive resilience quantification [14] directly. There are also parallel advances in methods driven by detailed outage data in distribution systems such as [15–17].

7.3 Transmission Utility Data

The New York Independent System Operator (NYISO) is the organization that manages New York State's electric grid and wholesale electric marketplace [18]. Detailed power grid outage data can be publicly accessed from the NYISO website [18]. The outage data on the website span from July 2002 to the present. For this chapter, we use twelve years of these outage data from November 2008 to November 2020.

NYISO uses data collection methods that check the system's current status every 5 minutes and record the status in a database. This 5-minute granularity of recording interval results in about 35 000 records per day and 12.6 million per year for each data type. The two data types that we are interested in using are the actual real-time outages and the scheduled outages. The real-time actual outage data records the current status of all outages present in the system at the time of checkpoint, including the timestamp, part identification (PTID), equipment name, and the outage date/time as shown in Table 7.1.

The timestamp is the checkpoint of the date and time at which the system recorded the information. Part identification (PTID) is a unique numerical tag identifying each system component. The equipment name for a transmission line identifies the names of the sending and receiving buses and the rated voltage; for example, N.SIMONE-COLTRANE_138_361. The equipment name for a transformer identifies the substation. Table 7.1 also shows outages of filter capacitors and circuit breakers. Note that the outage date/time is when the component went out, which is different from the timestamp. Although the data is public, we follow good practice in anonymizing the substation names in Tables 7.1 and 7.2.

Table 7.1: Real-Time Actual Outage Data

Timestamp	PTID	Equipment Name	Outage Date/Time
4/5/2021 2:22	25312	NRTHSIMONE_138N_138E_PAR 1	1/15/2018 10:15
4/5/2021 2:22	25126	WYNTON_120_SVC_CLC1	1/26/2018 10:15
4/5/2021 2:22	25913	DELFAYO_120KV_CAP_GC2 FILTER	1/15/2018 10:15
4/5/2021 2:22	25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29
4/5/2021 2:22	25908	BRADFORD345KV_8_____CB	1/26/2018 10:15
4/5/2021 2:22	25116	ELLIS_DC_GC1	3/12/2018 00:59
4/5/2021 2:22	25916	E.FITZGERALD-DAVIS_345_31	10/20/2020 17:09
4/5/2021 2:22	25917	GILLESPIE-ELLINGTON_345_30	1/25/2018 10:15
4/5/2021 2:22	25904	N.SIMONE-HOLIDAY_138_465	1/16/2020 10:15
4/5/2021 2:22	25905	SIMONE-N.SIMONE_C_115_3-VI	1/15/2018 4:13
4/5/2021 2:22	25937	WYNTON_120KV_CAP_GC1 FILTER	3/20/2018 00:15
4/5/2021 2:22	25921	MARSALIS 345KV_1500-A_____CB	4/23/2019 1:25
4/5/2021 2:22	25927	MARSALIS 345KV_77-2X_____CB	4/23/2019 10:13
4/5/2021 2:27	25912	DELFAYO120KV_120-101_____CB	1/15/2018 10:15
4/5/2021 2:27	25312	NRTHSIMONE_138N_138E_PAR 1	1/15/2018 10:15
4/5/2021 2:27	25126	WYNTON_120_SVC_CLC1	1/26/2018 10:15
4/5/2021 2:27	25913	DELFAYO_120KV_CAP_GC2 FILTER	1/15/2018 10:15
4/5/2021 2:27	25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29
⋮	⋮	⋮	⋮

The real-time scheduled data record the outages that are scheduled to occur for operational or maintenance reasons. The real-time scheduled data include the timestamp, PTID, equipment name, scheduled out date/time, and scheduled in date/time as shown in Table 7.2. The definition of the timestamp, PTID, and equipment name is the same as in the actual real-time data. The in date/time is the date and time that the component is scheduled to be re-energized. Commonly, scheduled outages are rescheduled.

Table 7.2: Real-Time Scheduled Outage Data

Timestamp	PTID	Equipment Name	Out Date/Time	In Date/Time
4/5/2021 2:22	25312	NRTHSIMONE_138N_138E_PAR 1	1/15/2018 10:15	12/6/2021 10:59
4/5/2021 2:22	25126	WYNTON_120_SVC_CLC1	1/26/2018 10:15	5/12/2022 10:15
4/5/2021 2:22	25913	DELFAYO_120KV_CAP_GC2 FILTER	1/15/2018 10:15	4/23/2021 2:22
4/5/2021 2:22	25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29	10/20/2021 12:59
4/5/2021 2:22	25908	BRADFORD345KV_8_____CB	1/26/2018 10:15	4/13/2023 4:59
4/5/2021 2:22	25116	ELLIS_DC_GC1	3/12/2018 00:59	1/13/2025 00:59
4/5/2021 2:22	25916	E.FITZGERALD-DAVIS_345_31	10/20/2020 17:09	1/25/2025 1:59
4/5/2021 2:22	25917	GILLESPIE-ELLINGTON_345_30	1/25/2018 10:15	2/1/2025 0:59
4/5/2021 2:22	25904	N.SIMONE-HOLIDAY_138_465	1/16/2020 10:15	2/1/2025 23:00
4/5/2021 2:22	25905	SIMONE-N.SIMONE_C_115_3-VI	1/15/2018 4:13	12/6/2021 3:45
4/5/2021 2:22	25937	WYNTON_120KV_CAP_GC1 FILTER	3/20/2018 00:15	10/13/2023 0:59
4/5/2021 2:22	25921	MARSALIS 345KV_1500-A_____CB	4/23/2019 1:25	3/28/2023 7:45
4/5/2021 2:22	25927	MARSALIS 345KV_77-2X_____CB	4/23/2019 10:13	3/20/2021 1:15
4/5/2021 2:27	25912	DELFAYO120KV_120-101_____CB	1/15/2018 10:15	5/26/2021 10:30
4/5/2021 2:27	25312	NRTHSIMONE_138N_138E_PAR 1	1/15/2018 10:15	12/6/2021 10:59
4/5/2021 2:27	25126	WYNTON_120_SVC_CLC1	1/26/2018 10:15	5/12/2022 10:15
4/5/2021 2:27	25913	DELFAYO_120KV_CAP_GC2 FILTER	1/15/2018 10:15	4/23/2021 2:22
4/5/2021 2:27	25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29	10/20/2021 12:59
⋮	⋮	⋮	⋮	⋮

The intended use of this data is to show that a publicly access dataset can be used for reliability assessments on a transmission system. The real-time scheduled and real-time actual data files need to be combined into one file that shows a comprehensive account both files in order to perform assessments.

7.4 Data Processing

Although many utilities record detailed outage data that would provide significant inputs to reliability models and simulations, most are not readily forthcoming of sharing this data with

researchers. There are a limited number of publicly access outage data resources available, but some processing is required to make it viable. Public accessible outage data in its raw form is heterogeneous and voluminous, with detailed outage information of all components in a system. This data in its raw state is impractical to use as a direct input into a model for analysis, especially for transmission line outages. Processing public accessible outage data into a consolidated form will allow for components such as transmission lines and transformers to be identified for researchers to use.

This section describes the details of the processing that compresses and combines the actual real-time outages and the real-time scheduled outages into a single dataset. The processing compresses the data to make it manageable, identifies the automatic outages, and removes repeated data. It is inherent in converting from data recording the outage status every 5 minutes to a list of outages described once that large amounts of repetitive data must be deleted. All the processing is done using Mathematica to help mitigate the difficulties of handling mixed alphanumeric and date and time data.

7.4.1 Compression

The objective of the compression is to discard most of the real-time data that is not needed to make the file sizes more manageable. The source files are for each day from November 2008 to November 2020 (except that some days are missing in April and October 2010 and September 2016). Each month of daily source files is read and compressed as follows.

The real-time actual outage data is very repetitive as the outage is recorded every 5 minutes until it is restored. The real-time actual outage data is sorted according to PTID, then Equipment Name, then Outage Date/Time, and then Timestamp. Then the outages are grouped according to the same successive PTID, Equipment Name, and Outage Date/Time. Only the first and last of each group (with the minimum and maximum timestamp, respectively) are retained. This removes most of the repeated records for the same outage and compresses the real-time actual outage data.

The real-time scheduled outage data is very repetitive. The scheduled outage is recorded every 5 minutes until it happens, and the scheduled outages are frequently postponed to a later time. The timestamp is removed, and then duplicate records are discarded. Then only those outages that are not postponed are retained: the successive pairs of scheduled outages that either have different PTID or have Out Date/Times differing by more than 16 minutes are determined to be not postponed and are retained. This leaves a record of only the last time the outage was scheduled in a month and compresses the real-time scheduled outage data.

Finally, the monthly compressed data is combined into a single dataset and sorted according to Out Date/Time for each actual and scheduled real-time data.

7.4.2 Identifying automatic outages

One objective of data processing is to identify the automatic outages. This is done by noting the outages that occurred but were not scheduled. The automatic outages are those in the actual real-time outages but not in the real-time scheduled outages.

In detail, for each actual outage, the scheduled outage with the same PTID with scheduled Out Date/Time closest in time to the actual Out Date/Times is searched for. If there is no such scheduled outage, or the closest scheduled Out Date/Time is more than one hour different than the actual Out Date/Time, then the actual outage is identified as automatic. If the closest scheduled Out Date/Time is less than one hour different than the actual Out Date/Time, then the actual outage is identified as scheduled. The processing can now neglect the scheduled outage data and proceed with the actual outages identified as automatic or scheduled. We were unable to deduce usable component repair times from the data. The final step is to remove any remaining repeated records of the same outage. Any successive duplicated records of an outage with the same PTID, Equipment Name, and Out Date/Time are removed as in Table 7.3.

Table 7.3: Compressed Real-Time Outage Data

PTID	Equipment Name	Out Date/Time	In Date/Time	Outage Type
25312	NRTHSIMONE_138N_138E_PAR 1	1/15/2018 10:15	12/6/2021 10:59	Automatic
25126	WYNTON_120_SVC_CLC1	1/26/2018 10:15	5/12/2022 10:15	Scheduled
25913	DELFAYO_120KV_CAP_GC2 FILTER	1/15/2018 10:15	4/23/2021 2:22	Scheduled
25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29	10/20/2021 12:59	Automatic
25908	BRADFORD345KV_8_____CB	1/26/2018 10:15	4/13/2023 4:59	Scheduled
25116	ELLIS_DC_GC1	3/12/2018 00:59	1/13/2025 00:59	Automatic
25916	E.FITZGERALD-DAVIS_345_31	10/20/2020 17:09	1/25/2025 1:59	Automatic
25917	GILLESPIE-ELLINGTON_345_30	1/25/2018 10:15	2/1/2025 0:59	Scheduled
25904	N.SIMONE-HOLIDAY_138_465	1/16/2020 10:15	2/1/2025 23:00	Automatic
25905	SIMONE-N.SIMONE_C_115_3-VI	1/15/2018 4:13	12/6/2021 3:45	Scheduled
25937	WYNTON_120KV_CAP_GC1 FILTER	3/20/2018 00:15	10/13/2023 0:59	Scheduled
25921	MARSALIS 345KV_1500-A_____CB	4/23/2019 1:25	3/28/2023 7:45	Scheduled
25927	MARSALIS 345KV_77-2X_____CB	4/23/2019 10:13	3/20/2021 1:15	Scheduled
25912	DELFAYO120KV_120-101_____CB	1/15/2018 10:15	5/26/2021 10:30	Scheduled
⋮	⋮	⋮	⋮	⋮

7.4.3 Extracting transmission line outages

The transmission lines in the outage data have a standard format in their Equipment Name of two 8 character sending and receiving bus names separated by a hyphen, followed by the rated voltage and other information. It is straightforward to extract the transmission line outages by detecting this format (select Equipment Names with the 9th character a hyphen), as shown in Table 7.4.

Table 7.4: Real-Time Transmission Line Outage Data

PTID	Equipment Name	Out Date/Time	In Date/Time	Outage Type
25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29	10/20/2021 12:59	Automatic
25916	E.FITZGERALD-DAVIS_345_31	10/20/2020 17:09	1/25/2025 1:59	Automatic
25917	GILLESPIE-ELLINGTON_345_30	1/25/2018 10:15	2/1/2025 0:59	Scheduled
25904	N.SIMONE-HOLIDAY_138_465	1/16/2020 10:15	2/1/2025 23:00	Automatic
25905	SIMONE-N.SIMONE_C_115_3-VI	1/15/2018 4:13	12/6/2021 3:45	Scheduled
⋮	⋮	⋮	⋮	⋮

The results of processing the twelve years of data in 45 178 transmission line outages, comprising 9600 automatic line outages and 35 578 scheduled line outages.

7.4.4 Forming the network

The first step in forming the network is to clean the bus names. There can be slight variations in spaces, punctuation, or abbreviation that prevent the bus from being uniquely identified by its bus name that needs to be resolved.

After the bus names are cleaned, since almost all transmission lines have a planned or automatic outage in 12 years of observation, it is feasible to form the network from outage data simply by adding a link between the sending and receiving buses of each line that was outaged in the data, as explained in detail in [6]. A vital feature of the resulting network is that it is entirely compatible with the outage data in that, by construction, all the outaged lines can be located on the network. (Note that it can be difficult in practice to precisely relate the outages with other descriptions of the network.) There are 1192 buses in the cleaned bus data. Forming the network directly from the outage data yields a sizeable connected component as shown in Figure 7.1 of 1139 buses. The majority of the 53 buses are not in the large connected component are in portions of the grid outside New York state that are represented less comprehensively. 95.5% of the lines have voltage ratings ranging from 69 kV to 500 kV.

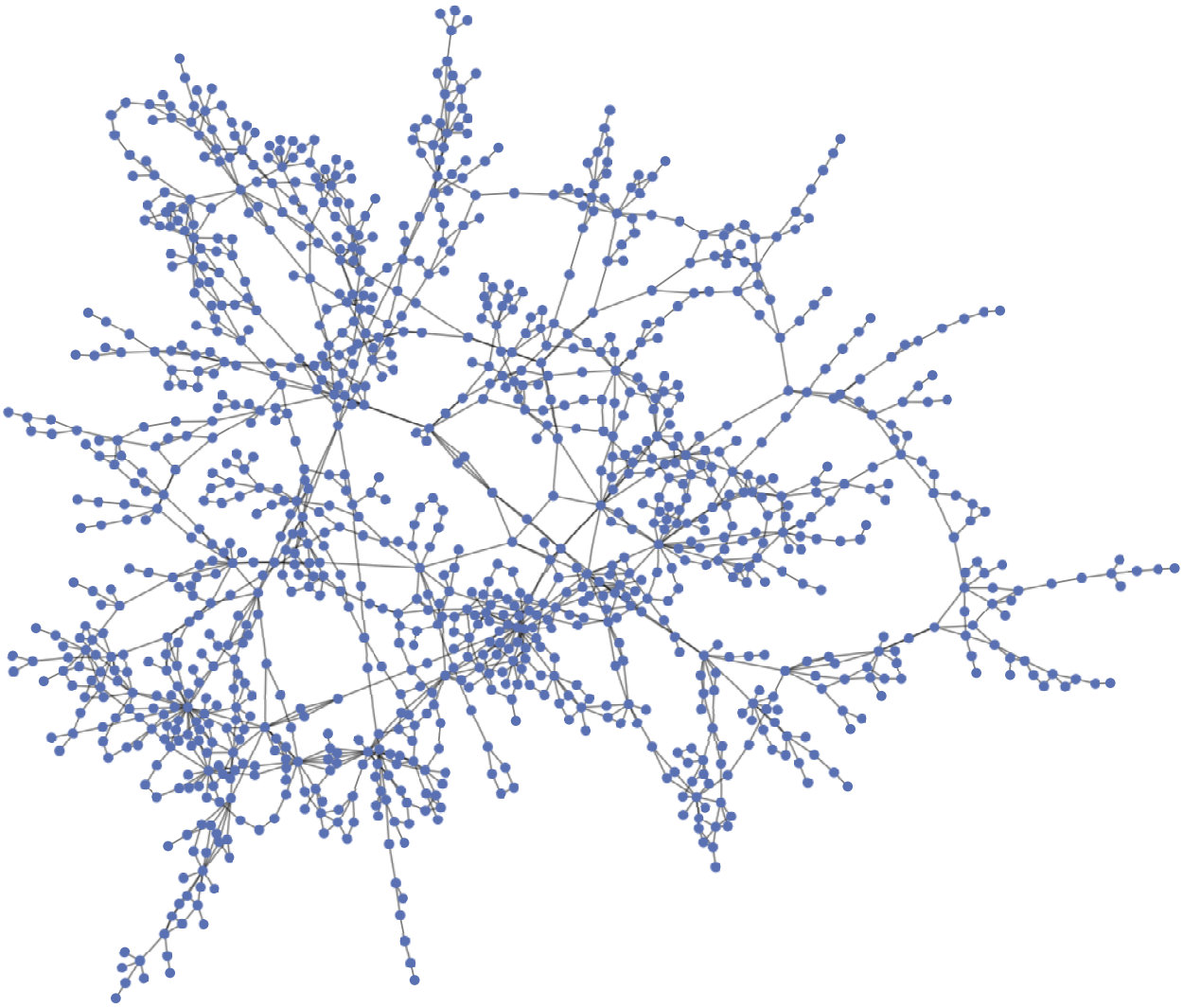

Figure 7.1: Network formed from the outage data.

7.5 Outage Statistics

This section shows some bulk statistics derived from the automatic transmission line outage data that describe how the line outages cascade on the network. The methods used to derive the statistics are the same as in [5, 6, 19], where they were used to process the detailed outage data from BPA. The numerical values of the plotted data are given in the appendix to facilitate researchers making qualitative comparisons of the results with other simulated or real data.

The outage data is grouped into cascades and generations based on the outage start time using the simple method described in [5].[1] An outage occurring more than one hour after the preceding outage is assumed to start a new cascade, and within each cascade a series of outages less than one minute apart are grouped into the same generation. Thus each cascade consists of a series of generations, with each generation containing one or more line outages that occur closely spaced in time. For example, outages caused by protection within one minute are grouped together in the same generation. This processing produces 6687 cascades. Since the power system is generally resilient, 66% of cascades have only one outage, and 84% of cascades have only a single generation of outages that does not propagate further.

The initiating line outages are those in the first generation of outages. The probability distributions of the number of initiating line outages and the number of line outages in each cascade are shown in Figure 7.2, and the corresponding survival functions are shown in Figure 7.3.

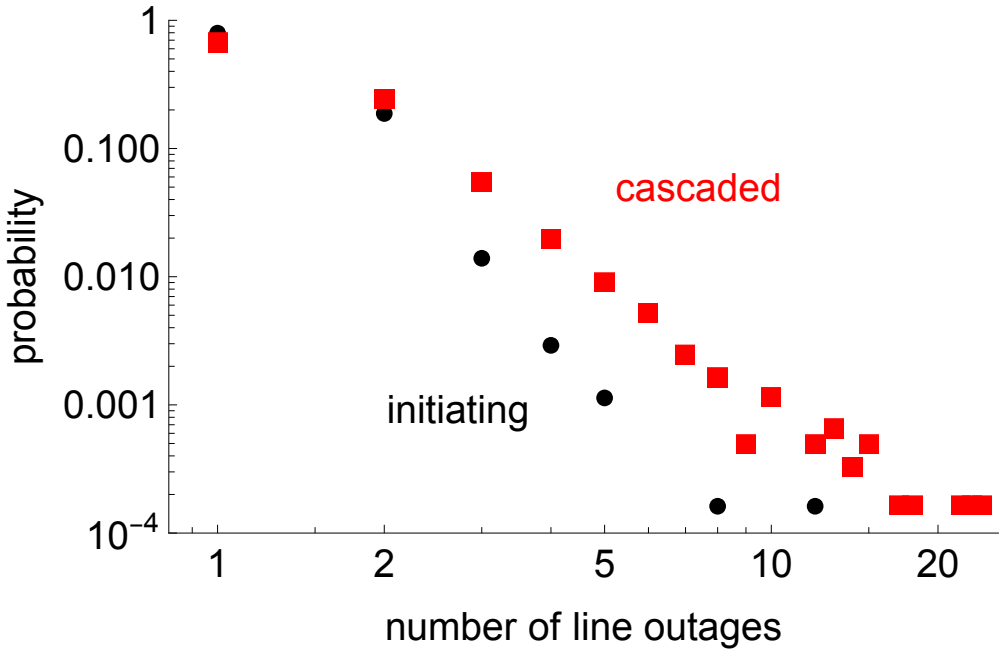

Figure 7.2: Probability distributions of the number of line outages in initiating and cascaded outages.

[1] Note that alternative ways of grouping outages into events are being developed [20].

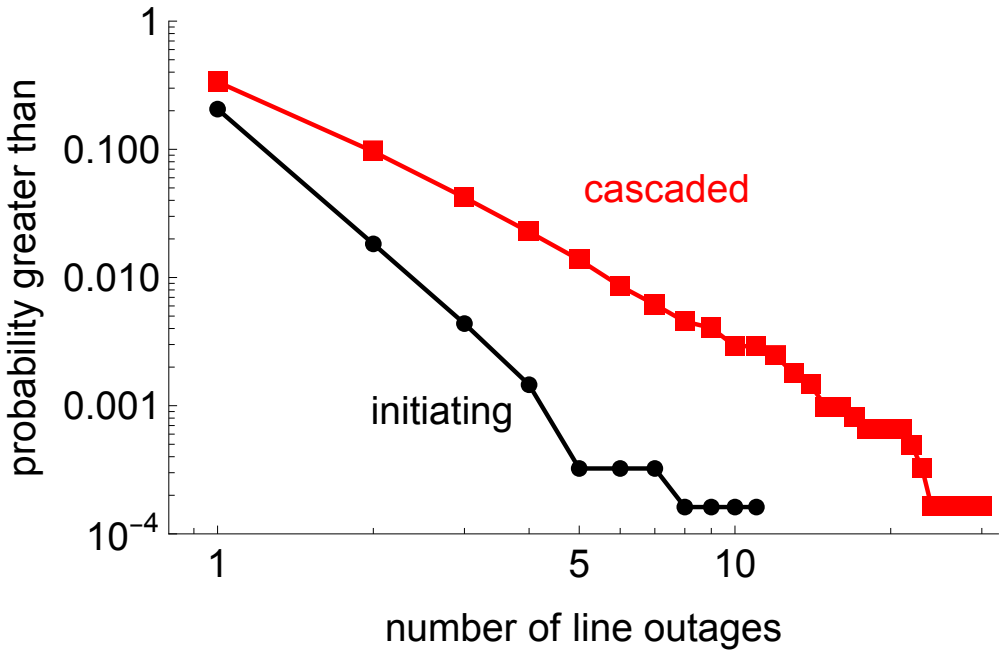

Figure 7.3: Survival functions of the number of line outages in initiating and cascaded outages.

Figures 7.2 and 7.3 show how cascading increases the number of line outages beyond the initiating outages. Note the heavy-tailed nature of the distributions, which is also seen in the analysis of BPA data in [5, Figure 1].

The propagation from generation k to generation $k+1$ in terms of the number of lines is defined as

$$\lambda(k) = \frac{\#\text{ lines out in generation } k+1}{\#\text{ lines out in generation } k}.$$

The line propagation in each generation is shown in Figure 7.4. The line propagation increases from a low value and then becomes more noisy for the higher generations due to the sparse data for the longer cascades. This general behavior is also seen in the analysis of BPA data in [5, Figure 3].

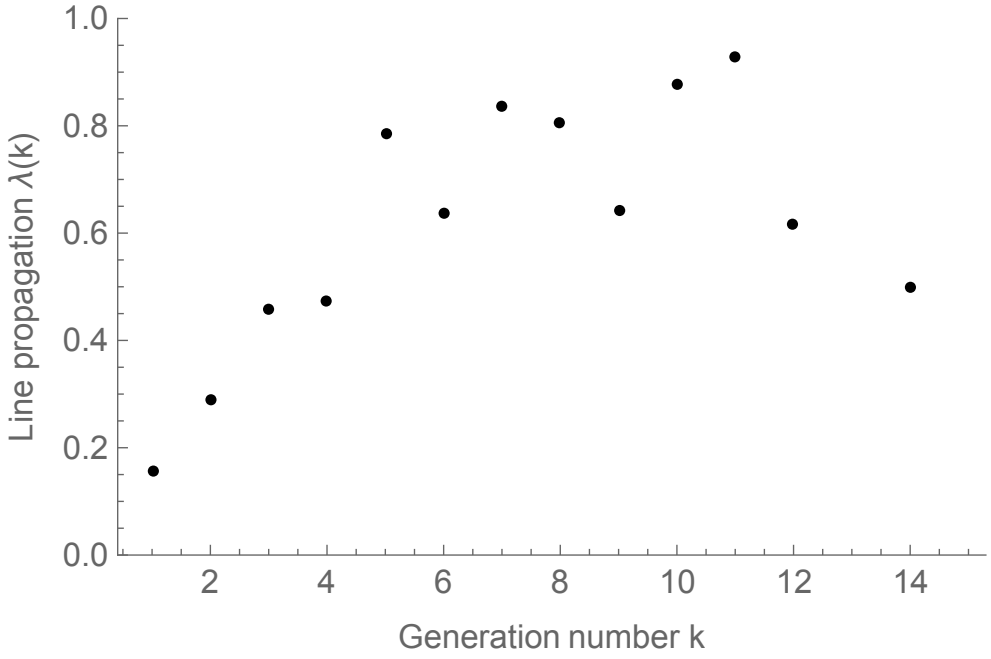

Figure 7.4: Line propagation $\lambda(k)$ as a function of generation number k.

A better way to measure propagation [19] uses the probability distribution of the number of generations in cascades or events as shown in Figure 7.5. The absolute value of the slope of the fitted red line in Figure 7.5 is the System Event Propagation Slope Index, or SEPSI, that is a single number describing the propagation of the generations [19, 20]. In this case SEPSI = 3.17. Figure 7.5 can be compared with the analysis of BPA data in [19, Figure 2] and with NERC data in [20, Figure 4]. However, a strict quantitative comparison with [20, Figure 4] is not appropriate because [20] uses a different grouping of outages into events or cascades than this chapter or [19].

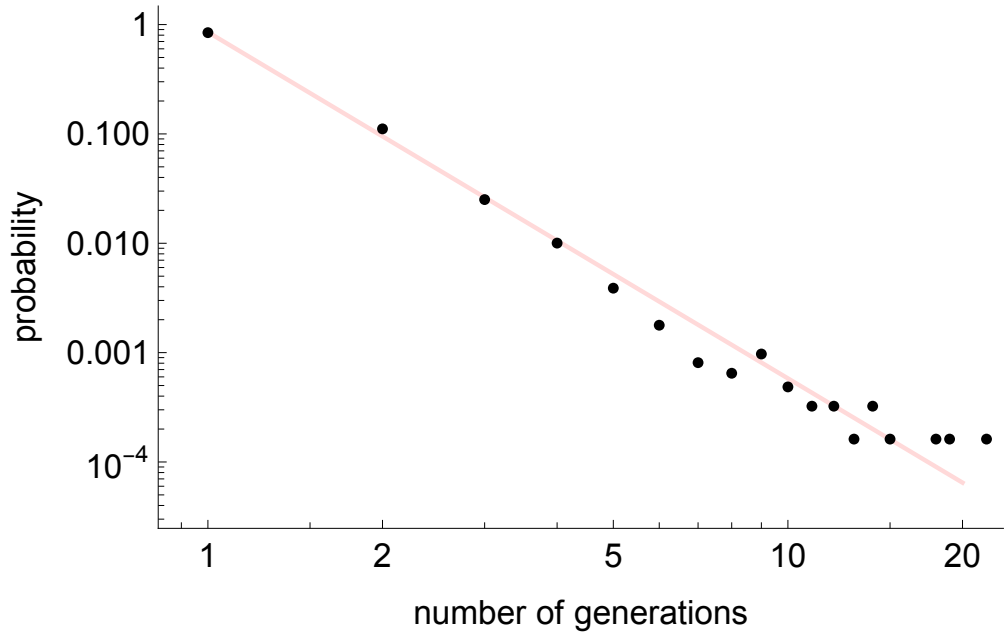

Figure 7.5: Distribution of number of generations in cascades.

The network distance between two lines can be measured as the number of "hops" on the network between the lines [6].[2] For example, the distance of line to itself is zero and the distance of a line to a neighboring line with at least one bus in common is one.

A better way to measure propagation [19] uses the probability distribution of the number of generations in cascades or events as shown in Figure 7.5. The absolute value of the slope of the fitted red line in Figure 7.5 is the System Event Propagation Slope Index, or SEPSI, that is a single number describing the propagation of the generations [19, 20]. In this case SEPSI = 3.17. Figure 7.5 can be compared with the analysis of BPA data in [19, Figure 2] and with NERC data in [20, Figure 4]. However, a strict quantitative comparison with [20, Figure 4] is not appropriate because [20] uses a different grouping of outages into events or cascades than this chapter or [19].

The network distance between two lines can be measured as the number of "hops" on the network between the lines [6].For example, the distance of line to itself is zero and the distance of a line to a neighboring line with at least one bus in common is one.

[2]More precisely, the network distance between lines L_i and L_j is defined as the minimum number of buses in a network path joining L_i to L_j.

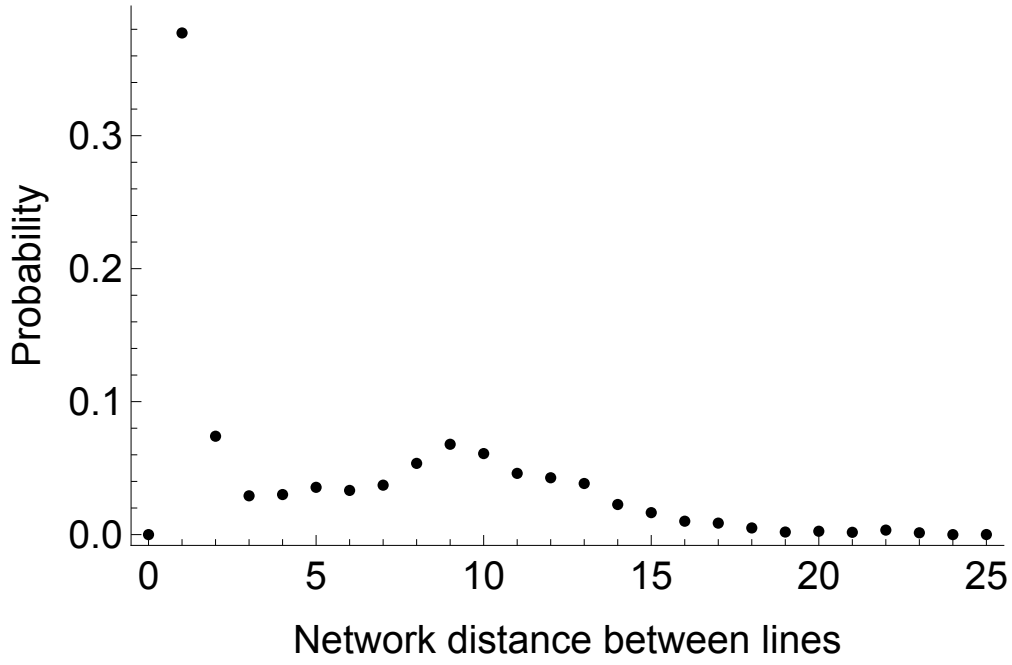

Figure 7.6: Distribution of network distances between random pairs of distinct lines in the same cascade.

7.6 Conclusions

The work in this chapter shows how to process data from a public website for a region of Northeast America that logged transmission grid outages every 5 minutes. We obtained a detailed list of component outages that occurred, with the outage times to the nearest minute, the component details, and whether the outage was scheduled or automatic. This is only the second such public source of detailed outage data for transmission grids to the authors' knowledge. We were unable to extract repair times from the data. The 12 years of processed data is sufficient to form a network on which the outages can be located. The detailed outage data is valuable to engineers and researchers, especially in studying the rarer dependencies between outages that occur in cascading and resilience events. The outage data is rich with possibilities, including the study of outages of a variety of equipment. To illustrate one of the uses of the data, we show how bulk statistics were obtained from the automatic transmission line outages in the data can be used to

quantify how cascading or resilience events propagate and spread. These statistics were formed similar to those in the other publicly available source of detailed outage data, showing that the main features of the previous work with this other source of data are reproduced in another region of North America. The statistics from this and the other public source are foundational in ensuring realism and validation of simulations and models of cascading and resilience.

7.7 References

[1] N. K. Carrington, I. Dobson, and Z. Wang, "Transmission grid outage statistics extracted from a web page logging outages in Northeast America," in *The 53rd North American Power Symposium (NAPS 2021*, 2021.

[2] "North American Electric Reliability Corporation." [Online]. Available: https://www.nerc.com/

[3] "Bonneville Power Administration transmission services operations & reliability website." [Online]. Available: https://transmission.bpa.gov/Business/Operations/Outages

[4] I. Dobson, N. Carrington, K. Zhou, Z. Wang, B. Carreras, and J. M. Reynolds Barredo, "Exploring cascading outages and weather via processing historic data," in *Hawaii International Conference on System Sciences 2018*, Big Island, Hawaii, USA, January 2018.

[5] I. Dobson, "Estimating the propagation and extent of cascading line outages from utility data with a branching process," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 2146–2155, November 2012.

[6] I. Dobson, B. A. Carreras, D. E. Newman, and J. M. Reynolds-Barredo, "Obtaining statistics of cascading line outages spreading in an electric transmission network from standard utility data," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4831–4841, November 2016.

[7] B. A. Carreras, D. E. Newman, I. Dobson, and N. S. Degala, "Validating OPA with wecc data," in *2013 46th Hawaii International Conference on System Sciences*. Maui, HI USA,: IEEE, January 2013, pp. 2197–2204.

[8] B. A. Carreras, J. M. Reynolds-Barredo, I. Dobson, and D. E. Newman, "Validating the opa cascading blackout model on a 19402 bus transmission network with both mesh and tree structures," in *2019 52nd Hawaii International Conference on System Sciences*, January 2019.

[9] J. Qi, "Utility outage data driven interaction networks for cascading failure analysis and mitigation," *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 1409–1418, March 2020.

[10] B. Gjorgiev, B. Li, and G. Sansavini, "Calibration of cascading failure simulation models for power system risk assessment," in *Safety and Reliability-Safe Societies in a Changing World-Proceedings of the 28th International European Safety and Reliability Conference, ESREL 2019*, September 2019.

[11] M. Noebels, R. Preece, and M. Panteli, "AC cascading failure model for resilience analysis in power networks," *IEEE Systems Journal*, December 2020.

[12] J. Bialek, E. Ciapessoni, D. Cirio, E. Cotilla-Sanchez, C. Dent, I. Dobson, P. Henneaux, P. Hines, J. Jardim, S. Miller *et al.*, "Benchmarking and validation of cascading failure analysis tools," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4887–4900, November 2016.

[13] P. Henneaux, E. Ciapessoni, D. Cirio, E. Cotilla-Sanchez, R. Diao, I. Dobson, A. Gaikwad, S. Miller, M. Papic, A. Pitto *et al.*, "Benchmarking quasi-steady state cascading outage analysis methodologies," in *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, June 2018, pp. 1–6.

[14] M. R. Kelly-Gorham, P. D. Hines, K. Zhou, and I. Dobson, "Using utility outage statistics to quantify improvements in bulk power system resilience," *Electric Power Systems Research*, vol. 189, p. 106676, December 2020.

[15] N. K. Carrington, I. Dobson, and Z. Wang, "Extracting resilience metrics from distribution utility data using outage and restore process statistics," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5814–5823, 2021.

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

[17] C. Ji, Y. Wei, H. Mei, J. Calzada, M. Carey, S. Church, T. Hayes, B. Nugent, G. Stella, M. Wallace *et al.*, "Large-scale data analysis of power grid resilience across multiple us service regions," *Nature Energy*, vol. 1, no. 5, pp. 1–8, April 2016.

[18] "New york independent system operator website." [Online]. Available: https://www.nyiso.com/

[19] I. Dobson, "Finding a Zipf distribution and cascading propagation metric in utility line outage data," *arXiv preprint arXiv:1808.08434*, August 2018.

[20] S. Ekisheva, R. Rieder, J. Norris, M. Lauby, and I. Dobson, "Impact of extreme weather on North American transmission system outages," Washington DC USA, July 2021.

CHAPTER 8. GENERAL CONCLUSION

8.1 Narrative of contributions

In this thesis, a process that automates the extraction of resilience curves from detailed utility outage data was developed. The detected resilience curves were grouped into three sizes based on the total number of outaged components. A conventional three-stage framework was applied to all detected curves in order to extract resilience metrics for a range of event sizes. Resilience metrics such as outage process duration and recovery rate were extracted from the data for each size group using resilience triangles. The mean, median, standard deviation, and the upper bound 95% confidence interval of the metrics was extracted for each group of events. As a result of applying the three-stage framework to real distribution utility data, undesirable outcomes and processing difficulties were observed since outages and restores overlapped in time. When an outage and restore overlap in time they cancel out each other during the accumulation that forms resilience curves, obscuring the structure of the event. The problem of overlapping outages and restores was solved by developing a way to systematically decompose resilience curves into recovery and outage processes. The separate processes are pure, consisting of only outage and only restores while keeping the same event duration. Mathematical formulas were derived from the restore and outage processes to extract resilience metrics as a function of the number of outages. This was done by fitting functions to the mean and standard deviation of parameters of the recovery and outage processes to smooth and interpolate the data, to give the statistics of resilience metrics.

A gamma distribution was fitted using the method of moments to estimate the variability of duration resilience metrics as a function of the number of outages. Currently, utilities have the ability to predict the number of outages of an anticipated or ongoing event. The estimates of the variability of duration resilience metrics based on the number of outages provide a way to

estimate a worst case variability of duration resilience metrics so that utilities can tell their customers when a blackout will end with a specified degree of confidence.

The resilience event risk to customers has been captured as a function of the number of outages. The analysis shows that risk decreases as event size increases.

Dispatcher cause codes are studied to identify that tree-limb-related causes are the primary outage cause, especially for the more homogeneous resilience events larger than 30 outages. Using daily climate data from NOAA with the utility data a relation between average wind speed by event size was obtained.

The origin and propagation of initial outages was identified by analyzing one large North American transmission utility's 14-year data set. With the inclusion of NOAA storm data, cascades were categorized by dispatcher cause codes as weather-related and non-weather related and the processing methods showed significantly greater propagation from the initial failures and a significantly higher outage rate.

A deployable tool was developed to analyze and clean distribution AMI data. The tool provides features such as k-means clustering to group customers into their rate class based on load consumption and gives a breakdown of load consumption per rate class. This work provides utilities with models that estimate resilience metrics and their variability from their own detailed utility data; and provides a standalone tool that models and analyzes customer loads using smart meter data.

The work also demonstrates how to condense and process voluminous logged transmission grid outage data from a public website of a transmission system utility in the Northeastern region of America into a valuable dataset. Detailed information such as component details and whether an outage was automatic or scheduled were identified within the data. Bulk statistics were obtained from the automatic transmission line outages identified from the processing to quantify the propagation and spread of cascading outages as an example of the value of the data.

8.2 Research Contributions Summary

The research contributions for this thesis are as follows:

- A process for systematically detecting resilience curves was developed. An event definition was developed to automatically find the beginning and the ending of the resilience curve in utility data. The points in between the beginning and the end of the resilience curve are the accumulated unrestored outages. The process of detecting event curves systematically was also applied to customer outage count utility data to find customer resilience curves.

- The three-stage framework was applied to the categorical groups of the resilience curves to divide each curve into hazard prevention, damage propagation, and restoration phases. Resilience triangles were used to approximate the damage propagation and restoration phases to extract the resilience metrics of recovery rate, outage rate, outage duration, and restoration duration.

- A mathematical method was developed to decompose resilience curves into two independent processes consisting of only outages or only restores from each event. Resilience metrics such as restore duration, customers hours not served, and outage rates were calculated from the processes. The average resilience metrics were determined as a function of event size by curve fitting. Then the mean and standardized deviations of the duration metrics were derived as a function of event size.

- The variability of the resilience metrics for duration was calculated using a 95% confidence interval. This enables utilities to predict when restoration will be completed with a high degree of confidence given the estimate of the event size.

- A risk function that considers customer impact as a cost was calculated. For this utility dataset the risk cost decreases as the event size increase.

- The outage and event causes were investigated using their dispatcher cause codes. Larger events are more homogeneous in cause. The dependence of event size on wind speed was

determined using NOAA data. The wind speed averaged over events increases with event size.

- A deployable software tool was developed to process smart meter data for small utilities. The tool allows small utilities to import and clean large amounts of smart meter data, perform simple analysis such as load contribution, load duration and customer classification and export all tabular data and graphics produced from classification for external use. A data management plan was utilized to assess the quality of utility data and to gain insightful information from real utility data.

8.3 Publications and Presentations

8.3.1 Publications

Published:

- N. K. Carrington, I. Dobson and Z. Wang, "Extracting Resilience Metrics From Distribution Utility Data Using Outage and Restore Process Statistics," in *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5814-5823, Nov. 2021, doi: 10.1109/TPWRS.2021.3074898.

- N. K. Carrington, I. Dobson, and Z. Wang, "Transmission grid outage statistics extracted from a web page logging outages in Northeast America," in *The 53rd North American Power Symposium (NAPS 2021)*, 2021. [1st Place Best Paper]

- N. K. Carrington, S. Ma, I. Dobson, and Z. Wang, "Extracting resilience statistics from utility data in distribution grids," in *2020 IEEE Power Energy Society General Meeting (PESGM)*, 2020, pp. 1–5.

- S. Ma, N. Carrington, A. Arif, and Z. Wang, "Resilience assessment of a self-healing distribution systems under extreme weather events," in *2019 IEEE Power Energy Society General Meeting (PESGM)*. IEEE, 2019.

- I. Dobson, N. Carrington, K. Zhou, Z. Wang, B. Carreras, and J. M. Reynolds Barredo, "Exploring cascading outages and weather via processing historic data," in *Hawaii International Conference on System Sciences 2018*, 09 2018.

- H. Sun, Z. Wang, J. Wang, Z. Huang, N. Carrington, and J. Liao, "Data-driven power outage detection by social sensors," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2516–2524,2016.

8.3.2 Presentations

Related presentations on this work is:

- Decentralized Voltage/VAR Control based on PV Inverters. The 2016 Institute of Electrical and Electronics Power &Energy Society Innovative Smart Grid Technologies (IEEE PES ISGT) North America Conference. September 2016.

- A tool for mining AMI data to model customer loads for small public power utilities. The 56^{th} Electric Power Research Center (EPRC) Annual Conference. Ames, IA. May 2018.

- Robust Real-Time Modeling of Distribution Systems with Data-Driven Grid-Wise Observably. Ames, IA. ECIAC Meeting April 2019.

- DEED Webinar: Managing Your AMI Data. American Public Power Association The Academy (APPA). Ames, IA. February 2019

- Resilience Assessment of Self-healing Distribution Systems Under Extreme Weather Events. The 2019 IEEE Power & Energy Society General Meeting (PESGM). Atlanta, Georgia. August 2019.

- A tool for mining AMI data to model customer loads for small public power utilities. The 2019 American Public Power Association Customer Connections Conference . New Orleans, LA. October 2019.

- Extracting Resilience Statistics from Utility Data in Distribution Grids. The 2020 IEEE Power Energy Society General Meeting (PESGM). Montreal,Canada. August 2020.

- Extracting resilience metrics from utility data. Electric Power Research Center (EPRC) Winter Updates Meeting. Ames,IA. December 2020.
- Mining Your AMI Data. The Missouri Public Utility Alliance (MPUA) Annual Conference. Columbia, MO. October, 2021
- Transmission grid outage statistics extracted from a web page logging outages in Northeast America. The 53^{rd} North American Power Symposium (NAPS 2021). College Station, TX. November 2021.
- DEED Webinar: AMI Insights for All American Public Power Association The Academy (APPA). Ames, IA. November 2021

Other presentations:

- Decentralized Voltage/VAR Control based on PV Inverters. The 2016 Institute of Electrical and Electronics Power &Energy Society Innovative Smart Grid Technologies (IEEE PES ISGT) North America Conference. September 2016.
- Robust Real-Time Modeling of Distribution Systems with Data-Driven Grid-Wise Observablity. Ames, IA. ECIAC Meeting April 2019.