BAYESIAN AND STOCHASTIC NETWORK METHODS TO MODEL AND OPTIMIZE THE

RESILIENCE OF CRITICAL INFRASTRUCTURE SYSTEMS

By

Jinzhu Yu

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Interdisciplinary Studies: Systems Engineering

December 12, 2020

Nashville, Tennessee

Approved:

Professor Hiba Baroud

Professor Dan Work

Professor Matthew Shotwell

Professor Sankaran Mahadevan

To my parents, Mr. Yuangen Yu and Mrs. Jingmei Li.

ACKNOWLEDGMENTS

Thushara, William, Eric, Yulin, Yingxiao, Kyle, Abhinav, Pranav, Zhen Hu, Dan Ao, Oliver, Xinyue, Zhiliang, Peng, Chen, Fangfei, Siyuan, Zimu, Ruize, Xiaoyu, Chengzhi, Jack, Sarah, Kofi, Xiangming, Drew, Yuanzhe, Zhangxin, Li, Ruoyu, Chunlei, Fei Gao, Bowen, Yue Hu, Yanbing, Yanqing, V.K., B.K., Devika, Lea, Charles, and so on. They made my life at Vandy so memorable and helped me grow into a better person in one way or another. Some of them also generously helped me with my research through many technical discussions, especially during my first year here.

I want to thank my family for their constant love and support over the years, which helps me to stay happy and grateful and has been a constant source of motivation for me to work hard.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

xi

Chapter 1

Introduction

1.1 Motivation

Critical infrastructure systems such as power grids, water distribution systems, telecommunication systems, and multi-modal transportation systems provide essential services to society [7, 8, 9]. However, these systems are rapidly deteriorating due to aging infrastructure and climate-induced disasters. On average, the infrastructure systems of the United States scored a D^+ in the evaluation by the American Society of Engineers in 2017 [10]. Driven by the development of social economy and advances in technology, the infrastructure systems are becoming increasingly interdependent, making them more vulnerable to perturbations due to the propagation of *cascading failures* [11, 12, 13]. As such, the need to improve the security and resilience of infrastructure systems has been increasingly emphasized by the U.S. government [14].

To this end, an important first step is to quantify the performance of infrastructure systems before and after a disruptive event in order to improve their resilience, i.e., reduce vulnerability and enhance recoverability, which are two of the most critical aspects of system resilience (Fig. 1.1). However, several challenges emerge in evaluating the vulnerability and estimating the recovery state of infrastructure systems.

First, historical data about component fragility under different hazards are often lacking due to the rare occurrence of extreme hazardous events, such as earthquakes and severe storms. Second, the ability to make an accurate estimation of important parameters for describing the recovery state is often hindered by the lack of empirical data about network recovery state, thereby leading to different levels of uncertainty around the damage and recovery states at the component and system levels. Neglect or improper characterization of the uncertainty can result in under- or over-estimation of systemic performance

as the metrics cannot be obtained within an acceptable level of fidelity or estimated accurately [15, 16, 17]. Therefore, it is critical to develop models that properly characterize such uncertainty to evaluate component vulnerability and network recovery state under different types of hazards and given limited data. Third, while the topology of certain individual real-world infrastructure systems may be available, the data about the interdependencies across the infrastructure systems are typically not available. This is because 1) collecting such data requires the coordination of operators from multiple infrastructure systems that might not always be run by the same utility company, and 2) even if the data have been collected, utility companies may not share the data due to security concerns. The lack of such data leads to *epistemic uncertainty* [18] around the structure of the *interdependent infrastructure systems* (IISs). As such, interdependent links across the infrastructure networks need to be first evaluated, before the performance assessment of IISs. Finally, the vulnerability and recoverability of critical infrastructure are dependent on the allocation of resources to protect network components before a disruption and schedule repair activities after a disruption. Existing models that integrate protection and restoration have relied on deterministic optimization methods [19, 20, 21, 22], which only solve a few independent representative damage scenarios. This may lead to sub-optimal solutions due to *aleatory uncertainty* (natural variability [18]) about the disruption.

Table 1.1: Research objectives and their category according to different metrics

Component of resilience	Uncertainty	Task	Objective
Vulnerability	Component damage state	Estimation	Assess the serviceability of infrastructure systems
Vulnerability	Network configuration	Estimation	Model the interdependency of infrastructure systems
Recoverability	Network recovery state	Estimation	Improve the estimation accuracy of recovery state
Recoverability	N/A	Estimation	Improve the inference efficiency of the model for estimating recovery state
Vulnerability Recoverability	Hazard	Optimization	Optimize the integrated vulnerability reduction and recovery enhancement

This dissertation aims to address these challenges by developing and applying Bayesian and stochastic network methods to assess and optimize the resilience of infrastructure systems, i.e. reduce vulnerability and improve recoverability (Table 1.1). In this chapter, Section 1.2 briefly introduces the background concepts and methods involved in this dissertation. Section 1.3 details each objective of this dissertation. The organization of this dissertation is provided in Section 1.4.

1.2 Background

This section first introduces the concept of network resilience. Next, IISs are described because infrastructure systems are becoming increasingly interconnected. Because 1) infrastructure systems operate interdependently yet the data about their interdependence are lacking and 2) there exist various sources of uncertainty, enhancing the resilience of critical infrastructure systems requires a multidisciplinary approach, thus we then introduce the stochastic network methods used to assess and optimize resilience, including stochastic block model and two-stage stochastic programming.

1.2.1 Network Resilience

The resilience of infrastructure systems is a key element of this dissertation and is therefore introduced here. First introduced in ecology by Holling [23], the concept of resilience has triggered significant interest during the past decades in several other fields, ranging from psychology to engineering [24]. Within the engineering domain, multiple definitions of resilience have been proposed by different reputable organizations [25, 26, 27] as well as several scholars in the field of engineering resilience analysis [1, 28, 24]. In many of the definitions, resilience is similarly referred to as "the performance of a system of networks before, during, and after a disruptive event" [1, 29, 30]. This dissertation adopts this definition as it lends itself for convenient quantification.

Consider an engineering system with its time-dependent performance denoted by $\varphi(t)$.

Depending on the system under study, the metric for quantifying the performance can be network capacity, maximum flow, flow rate, or network connectivity [2]. As can be seen from Fig. 1.1, the time-dependent systemic performance can be divided into five stages: Initially, the system stays at the original stable state with performance level $\varphi(t_0)$. Stricken by the disruptive event e at time t_e, the system starts to degrade due to reduced or loss of functionality of components. The systemic performance deteriorates until the maximum loss is incurred at time t_d. Then the system enters the disrupted state in which the preparation activities start, such as identification of impaired components, transportation of materials and crew for restoration, etc. After a period of time, the system begins to recover from the disruptive state S_d. The recovery continues until time t_f when the system reaches a new stable state.
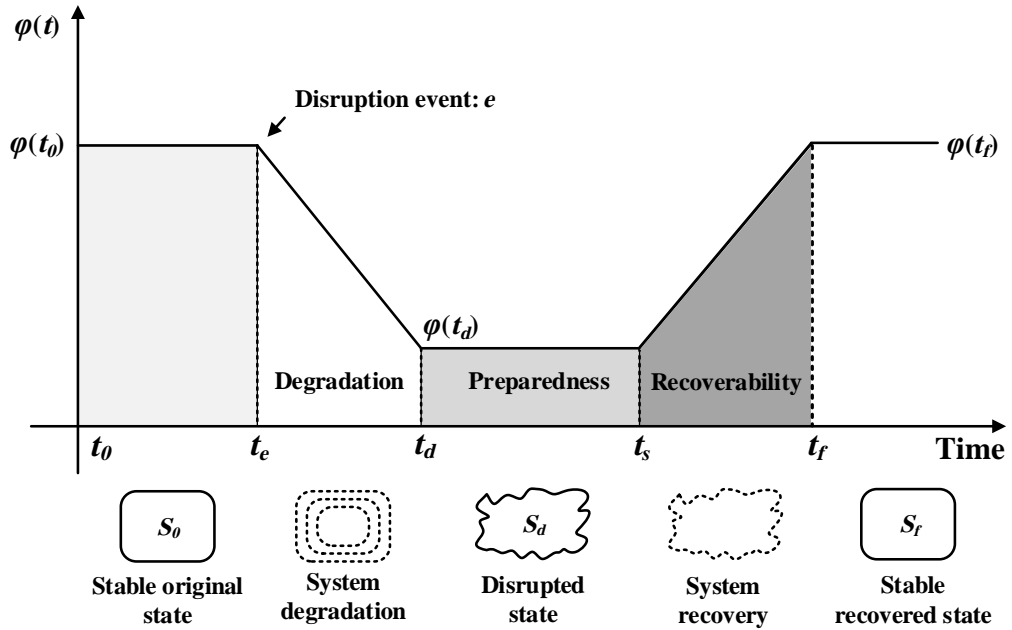
Figure 1.1: System performance over time (adapted from Ref. [1, 2])

Systemic resilience can be calculated as the ratio of recovered performance to the maximum loss [1]. Let R_t denote the resilience at time t during the recovery process. R_t can be computed as

$$R_t = \frac{\varphi(t|e) - \varphi(t_s|e)}{\varphi(t_0|e) - \varphi(t_s|e)} \tag{1.1}$$

4

where $t \in \left(t_s, t_f\right)$; φ represent the measure of systems performance or functionality level. Note that $R_t \in [0, 1]$ with 1 indicating that the system has been fully recovered from the disruptive event to the original state.

However, when the system is disrupted by some external disturbances or due to the malfunction or failure of internal components, it is difficult to describe the state of the system with mathematical and physical equations due to complex system dynamics or the lack of data [31].

1.2.2 Interdependent Infrastructure Networks

Interdependent infrastructure networks (i.e., interdependent infrastructure systems) are comprised of multiple individual networks connected by interdependency links. For example, the power network requires gas input from the gas network to generate electricity, the gas compressors depend on electricity from the power grid to transport gas; pumping stations in the water distribution network rely on electricity to lift water and power generators need clean water for cooling purposes (Fig. 1.2). This dissertation primarily considers those with two individual networks, however, the approach can be extended to accommodate additional networks.

Mathematically, $G = (G_1, G_2, E_{12}, E_{21})$ where $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ represent the individual networks and E_{12} and E_{21} represent the interdependency links from G_1 to G_2 and those from G_2 to G_1, respectively. The number of nodes and the number of links in G are denoted by $|V|$ and $|E|$ respectively. Each of the G_1 and G_2 is comprised of multiple supply nodes (S), transmission nodes (T), demand nodes (D), and the links that connect them (Fig. 1.3).

The connectivity of a graph can be encoded using the *adjacency matrix* where rows and columns are labeled by nodes. Let $A_{|V| \times |V|}$ represent the adjacency matrix of the graph. If there is a link between node i and node j, $A_{ij} = 1$, and $A_{ij} = 0$ otherwise. Since self-edges are not possible, we assume $A_{ii} = 0$, which is a valid assumption for the network
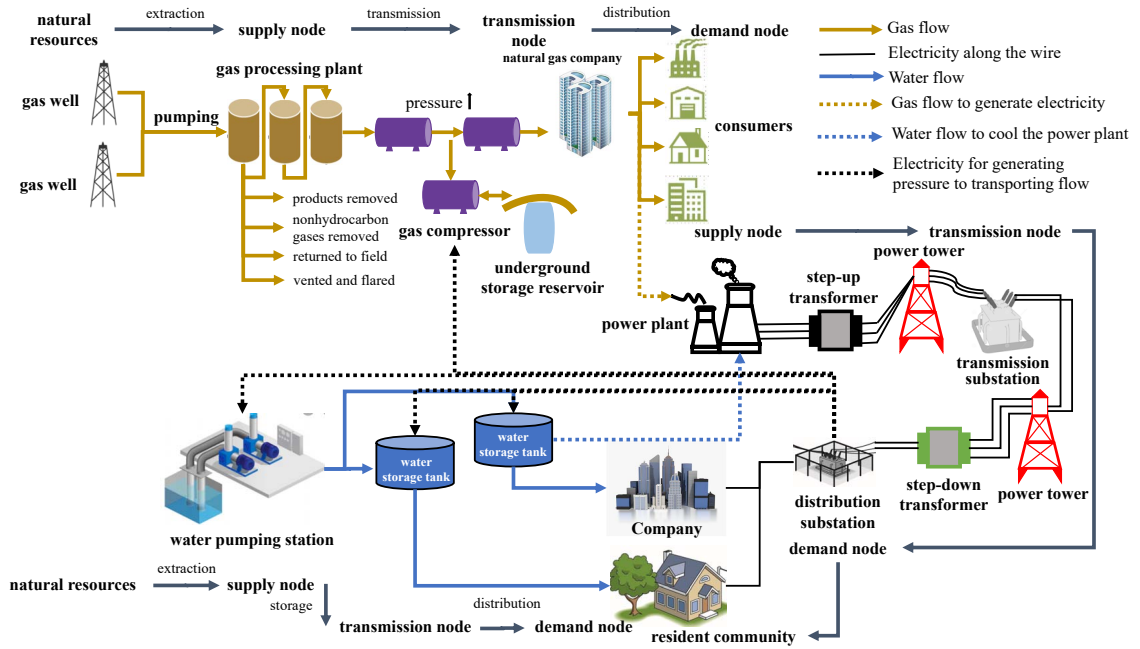
Figure 1.2: Schematic of interdependent power-gas-water systems [3]
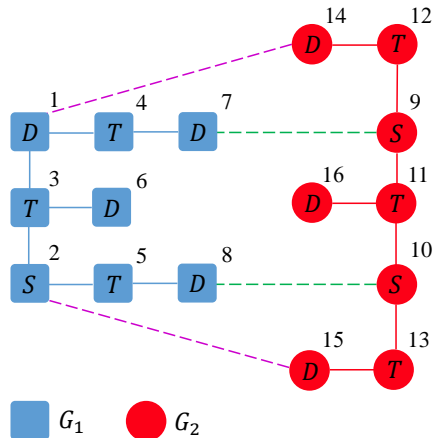
Figure 1.3: A schematic of abstract interdependent networks (Dash lines represent interdependency links. Nodes labeled with S, T, D represent supply nodes, transmission nodes, and demand nodes, respectively.)

representing infrastructure systems. For the undirected graphs considered in this study, the adjacency matrix is symmetric. As an illustrative example, the adjacency matrix for the interdependent networks in Fig. 1.3 is shown in Fig. 1.4.
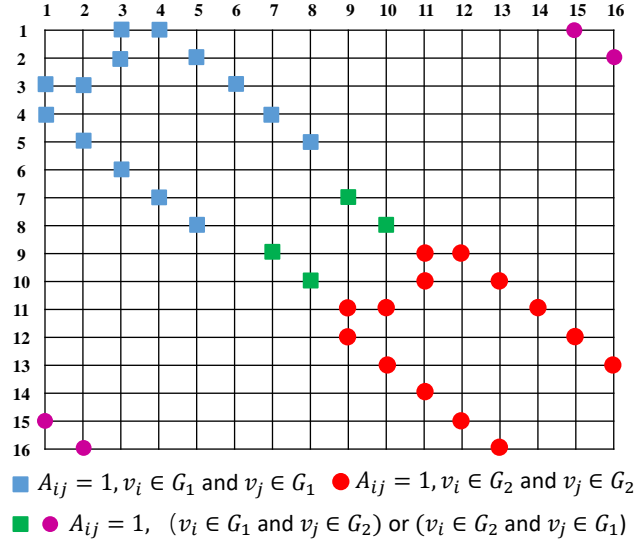
Figure 1.4: Adjacency matrix for the interdependent networks in Fig. 1.3

1.2.3 Stochastic Block Model

The *Stochastic block model* (SBM) is a probabilistic graphical model for describing and analyzing the structure of a network [32]. Pioneered by Holland et al. [33], SBM has been widely used in community detection in social networks [34, 35, 36]. In SBM, the nodes in a graph are divided into different blocks based on the class membership of the nodes. The probability of a link between two nodes depends on the blocks (class) to which the nodes belong. Let $z = [z_1, z_2, ..., z_{|V|}]^T$ represent the class membership vector and θ represent the *block probability matrix* [37], the matrix of probabilities of forming edges between blocks and within a block (Fig. 1.5).

Formally, SBM is defined as follows [33, 37]:

Definition 1.2.1 (SBM). *A is generated according to an SBM with respect to z if and only if: 1. $\forall i \neq j$, A_{ij} are statistically independent. 2. $\forall i \neq j$ and $i' \neq j'$ with $z_i = z_i'$ and $z_j = z_j'$, A_{ij} and A_{ij} are identically distributed.*

Given Definition 1.2.1, the block probability matrix, and the adjacency matrix, the probability of a link connecting node i in block a and j in block b can be given by Eq. (1.2).

7

Stochastic block model   Block probability matrix (*B*)
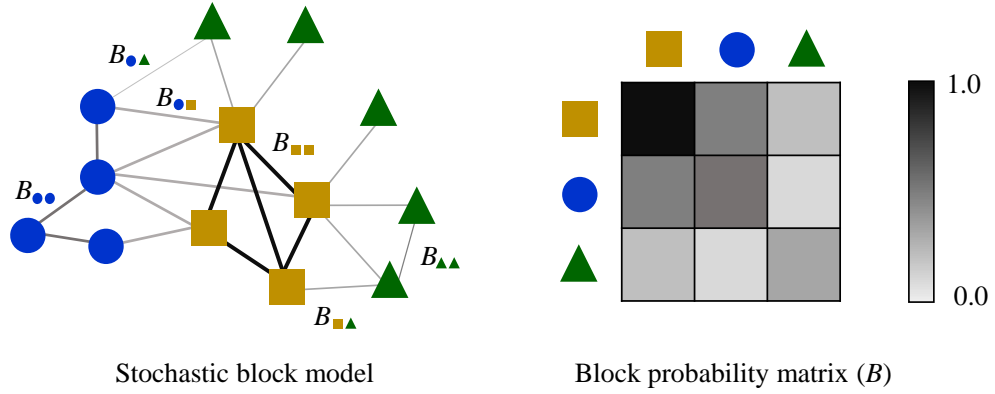
Figure 1.5: An example of SBM with the corresponding block probability matrix

$$P\left(A_{ij} = 1 | z_i = a, z_j = b\right) = \theta_{ab} \tag{1.2}$$

In Eq. (1.2), θ_{ab} represents the probability of forming a link between block *a* and block *b*. Equivalently, $A_{ij}|z_i = a, z_j = b \sim$ Bernoulli(θ_{ab}), meaning that entries of **A** can be modeled as statistically independent Bernoulli random variables [38]. Depending on the available data set, SBM can be applied to *a priori* setting where the partition of nodes are predefined and *a posteriori* setting where the block partition is uncertain [37]. In this study, SBM is used in the *a priori* setting.

The SBM offers several computational and application advantages such as the ability to estimate missing links based on incomplete data [39], the integration of statistical and network properties, and the flexibility in the analysis of stochastic interdependent links. However, this model does not consider the heterogeneity of nodes besides their block membership. In order to account for such heterogeneity, SBM is modified to give a probabilistic estimate of the presence of interdependency links E_{12} and E_{21}. The estimation of the probability of these links can be considered as a regression problem in which node attributes are the predictors and the probability of A_{ij} is the response variable. For example, suppose the class membership and the distance between nodes are used as the predictors, the regression model is shown in Eq. (1.3).

8

$$P\left(A_{ij} = 1 | z_1, z_2\right) = f\left(z_1, z_2, d_{v_i v_j}\right) \tag{1.3}$$

In Eq. (1.3), $d_{v_i v_j}$ represents the distance between v_i and v_j. Once the data on other nodal attributes are available, Eq. (1.3) can be modified to include more predictors, and the model form can be identified by the use of statistical methods for model selection.

1.2.4 Two-stage Stochastic Programming

Stochastic programming is an optimization approach to deal with decision making problems under uncertainty where the constraints or objective function are uncertain [40]. The uncertainty is typically characterized by a distribution that is known or can be estimated [41]. This dissertation involves the two-stage stochastic programming to model the problem, in which the decision makers take some actions before the occurrence of a random event, then the second-stage decisions (recourse decisions) are made to reverse the adverse impact of the random event [41]. A two-stage stochastic program seeks to minimize the cost function of the first-stage decisions $x \in \mathbb{R}^{n_1}$ and the expected cost associated with the second stage decisions $y \in \mathbb{R}^{n_2}$. The two decision epochs are divided by an random event, which is modeled using a random vector $\tilde{\omega}$ [42]. The optimal policy is a single first-stage decision and the second stage decisions to take in response to each realization of the random event [41].

A general two-stage stochastic program is of the following form:

$$\min_{x} \quad c^T x + \mathbb{E}\left[Q\left(x, \omega\right)\right] \tag{1.4a}$$

$$\text{s.t.} \quad Ax = b \tag{1.4b}$$

$$x \geq 0 \tag{1.4c}$$

where $c \in \mathbb{R}^{n_1}$, $b \in \mathbb{R}^{m_1}$, and $A \in \mathbb{R}^{m_1 \times n_1}$ are the known deterministic first-stage vectors and matrices, respectively. $\mathbb{E}\left[\cdot\right]$ represents the expectation over $\tilde{\omega}$. $Q\left(x, \tilde{\omega}\right)$ is the optimal

value of the second-stage problem, which is also called the recourse function. For a specific realization of the random event ω of $\tilde{\omega}$, $Q(x, \omega)$ is defined as:

$$Q(x, \omega) := \min_{y} \quad d(\omega)^T y \tag{1.5a}$$

$$\text{s.t.} \quad T(\omega)x + W(\omega)y = h(\omega) \tag{1.5b}$$

$$y \geq 0 \tag{1.5c}$$

where $W \in \mathbb{R}^{m_2 \times n_2}$, $T \in \mathbb{R}^{m_2 \times n_2}$, $d(\omega) \in \mathbb{R}^{n_2}$, and $h \in \mathbb{R}^{m_2}$ are given.

To solve the two-stage program numerically, it is assumed that $\tilde{\omega}$ has a finite number of realizations/scenarios. Let the probability of each scenario be $p_k(k = 1, ..., |\Omega|)$ and generate a copy of the second-stage decision variables for each scenario $\omega \in \Omega$ [42], the equivalent monolithic representation of the two-stage program above, often referred to as the *deterministic equivalent* or *extensive formulation*, is be given by

$$\min \quad c^T x \quad + \quad \sum_{k=1}^{|\Omega|} p_k d(\omega_k) y(\omega_k) \tag{1.6a}$$

$$\text{s.t.} \quad Ax \geq b \tag{1.6b}$$

$$T(\omega_1)x \quad + \quad W(\omega_1)y(\omega_1) \quad = \quad h(\omega_1) \tag{1.6c}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots$$

$$T(\omega_{|\Omega|})x \quad + \quad W(\omega_{|\Omega|})y(\omega_{|\Omega|}) \quad = \quad h(\omega_{|\Omega|}) \tag{1.6d}$$

$$x \geq 0, \qquad\qquad y(\omega) \geq 0, \qquad\qquad \omega \in \Omega \tag{1.6e}$$

This formulation has a nice structure because once the first-stage decision variables (complicating variables) are fixed, the problem naturally decomposes into $|\Omega|$ independent small sub-problems [43], rendering the problem easier to solve.

1.3   Research Objectives

This dissertation aims to assess and enhance the resilience of infrastructure systems under different sources of uncertainty by developing and applying Bayesian methods and stochastic network models. To achieve this goal, the research aims to accomplish the following objectives.

1. Assess the serviceability of infrastructure networks under uncertainty about the component damaged state in order to reduce the vulnerability against disruptions. A hierarchical Bayesian updating approach is proposed to evaluate the performance of network components under aleatory uncertainty and epistemic uncertainty. Then, the Bayesian updating approach is integrated into the assessment of network serviceability based on Monte Carlo simulation (MCS).

2. Improve the estimation accuracy of parameters describing the recoverability of infrastructure systems under epistemic uncertainty. To this end, a new class of models called hierarchical Bayesian kernel models (HBKM) is developed. The performance of HBKM is validated against the classical models using cross validation with historical data on community recovery from power outages.

3. Improve the inference efficiency of hierarchical Bayesian models for estimating the recovery state of infrastructure networks. The issue of computational efficiency arises in large-scale applications of hierarchical models due to the increased number of parameters Betancourt and Girolami [44]. Therefore, a scalable inference algorithm is developed to learn the hierarchical Bayesian model for count data, which are commonly encountered in calculating the recovery state of infrastructure systems. By providing fast yet competitively accurate inference of Bayesian hierarchical models, the proposed inference algorithm can inform the decision-making about restoration, especially in power grids where online estimates are required to enable the dynamic operations.

4. Model the interdependency across infrastructure networks under epistemic uncertainty, so that the vulnerability and recoverability of IISs can be evaluated with a higher level of accuracy. Nodal features of infrastructure systems are incorporated in the formulation of the modified stochastic block model to estimate the likelihood of interdependent links.

5. Optimize the integrated protection and restoration of IISs, i.e., reduce the vulnerability and increase recoverability from a systemic perspective, with considerations of uncertainty around disruptive events. The two-stage stochastic programming approach is leveraged to model the integrated problem of IISs. As the size of the problem may become too large for large-scale IISs, novel solution techniques are developed to obtain a good-quality solution within a reasonable amount of time.

The proposed research effort can help decision-makers in utility companies and emergency management agencies to improve the management of infrastructure systems. This research can eventually contribute to the secure and resilient operations of infrastructure systems, which are considered critical to social well-being, economic vitality, and national security [45].

1.4   Organization of the Dissertation

The rest of this dissertation is divided into several chapters, each addressing the research objectives.

Chapter 2 to Chapter 4 focus on the development and application of hierarchical Bayesian models as well as efficient Bayesian inference algorithms to improve the performance assessment of critical infrastructure systems. Specifically, Chapter 2 focuses on applying hierarchical Bayesian model to measure the performance of infrastructure systems under uncertainty. Component fragility metrics are estimated using the HBM and then integrated into a system serviceability assessment using Monte Carlo simulation and a shortest-path

algorithm. By using the Bayesian approach, the performance measures can be dynamically updated as more data becomes available. Estimating the system serviceability requires information about the component vulnerability and recoverability. However, making predictions at the component level presents challenges due to data scarcity. In order to address these challenges, Chapter 3 presents a novel class of models, hierarchical Bayesian kernel model (HBKM), to estimate the recovery rate of infrastructure networks in the aftermath of disruptions under scarce data. The proposed method is tested using cross-validation and compared with two other models for validation, the hierarchical Bayesian regression model (HBRM) and the Poisson Generalized Linear Model (GLM). A case study focusing on the recovery from power outage in Shelby County, Tennessee after the most severe storms since 2007 is presented to illustrate the proposed approach. The inference efficiency of hierarchical Bayesian Poisson regression model (HBPRMs) for count data, which are frequently encountered in estimating the recovery state of infrastructure systems, is considered in Chapter 4. As the applications of HBPRMs to large-scale datasets calls for more efficient inference algorithms, a scalable approximate Gibbs sampler (AGS) to learn the HBPRMs while maintaining the inference accuracy is developed. Numerical experiments on multiple real datasets demonstrate the superior performance of AGS against the state-of-the-art sampling algorithm, especially in large datasets.

Critical infrastructure systems do not operate individually but are rather interconnected. Theses interactions are characterized by interdependencies which are bidirectional relationships between two systems. Therefore models need to characterize these interdepndencies when assessing the performance of infrastructure systems. However, data on interdependent links is not available. To enable a more holistic approach to the performance assessment of infrastructure systems, Chapter 5 focuses on modeling their interdependencies under uncertainty. The uncertain interdependency links are modeled using a modified SBM, in which the probability of links between individual systems are estimated by considering nodal features in addition to the block membership of nodes. Then the modified stochastic

block model is integrated into the resilience assessment of IISs. Two recovery strategies based on static and dynamic component importance ranking are developed and compared.

While component importance ranking provides insights towards a decision making process, a formal resource allocation model is needed to determine strategies for reducing vulnerability and improving recoverability. Such decisions about resource allocation for vulnerability reduction and restoration improvement are typically made sequentially [20]. The separation of vulnerability reduction and restoration improvement decisions can lead to sub-optimal resilience, Chapter 6 investigates how to optimize the integrated protection and restoration planning of IISs. With considerations to the uncertainty around disruptions, a two-stage mixed-integer stochastic program is proposed to tackle the integrated decision-making problem. Specifically, the first stage considers the decisions about which components to fortify, while the second stage considers decisions concerning the schedule of repair crews, network flow, restoration sequence of damaged components, etc. Since binary decision variables are present in the second-stage problem, the proposed two-stage stochastic program cannot be solved effectively using classical approaches, e.g., the *L-Shaped* method (a.k.a. Bender's decomposition). To address this challenge, novel solution algorithms are proposed by integrating different meta-heuristics into the iterative regularization method. Synthetic interdependent power-gas networks are employed to illustrate the proposed two-stage stochastic program and solution techniques.

In Chapter 7, contributions made in this dissertation are summarized and directions for future work are discussed.

Chapter 2

Hierarchical Bayesian Approach to Assessing the Serviceability of Infrastructure
Networks Under Uncertainty

2.1   Introduction

Critical infrastructure systems have grown in scale and complexity as a result of urbanization and economic development [46]. Critical infrastructure systems are prone to a wide range of natural (i.e., extreme weather events) and anthropogenic (i.e., man-made attacks) hazards that make them vulnerable [47]. The U.S. government has placed an emphasis on protecting critical infrastructure and improving its ability to withstand and recover from all hazards as it is considered paramount to national security [14]. The protection of critical infrastructure systems relies on risk-informed decision making to identify strategies for hardening, emergency response, and restoration. In order to inform such decisions, it is essential to provide a reliable assessment of critical infrastructure system-level performance under different types of hazards. The challenge lies in the availability of sufficient data to make accurate estimates of the system performance metrics which include the component functionality and system serviceability. The proposed method is founded in a hierarchical Bayesian approach and combined with simulation and network algorithms to address challenges of data scarcity and uncertainty in the estimation of infrastructure network performance after disasters. While the study in this chapter focuses primarily on assessing the serviceability of water distribution systems (WDSs) impacted by earthquakes and floods, the proposed framework can be applied to other types of infrastructure networks, such as gas networks, oil networks, and power grid, among others, as long as the physical laws in the respective network are appropriately taken into account.

A water distribution system, which provides clean drinking water and wastewater treatment services, is credited to be one of the most essential infrastructures, especially during

a disaster when reduced water supply disrupts the emergency response (e.g., fist-aid, fire-fighting) and residential and industrial activities, among others [48, 49]. Unfortunately, WDSs have been overlooked and undervalued in the U.S. during the past decade. In 2019, the American Society of Civil Engineers gave the drinking water and wastewater infrastructure in the United States a D and D^+, respectively, due to aging infrastructure and underinvestment. Given insufficient resources, at the current replacement rate of 0.5% per year, it would take 200 years to replace the entire drinking water system which is 100 years beyond the system design life [50, 51]. The underinvestment is reflected in the system performance where approximately 240,000 water main breaks occur per year, with research indicating that as much as 20% - 25% of distributed water is lost due to leakage [52]. In addition to the performance decrease due to aging and underinvestment, disasters can affect the performance of WDSs drastically. Among the natural hazards, earthquake and flood have more impact on the performance of critical infrastructures [53].

Prior studies on the probabilistic assessment of infrastructure performance during disasters have focused on either (i) the estimation of component functionality (e.g., pipe failures [54], water main breaks [55]) or (ii) simulating the system's response to hazard (e.g., connectivity-based network models [56]). The uncertainty of the failure estimation of components has rarely been considered and integrated directly into the system-level serviceability assessment. The study in this chapter formally takes into consideration uncertainty and develops an approach that incorporates a hierarchical Bayesian updating mechanism of network component fragility into the evaluation of systemic performance of WDSs under earthquake and flood using Monte Carlo simulation and network algorithms. More specifically, this chapter improves the state-of-the-art in assessing the performance of critical infrastructure systems by making the following contributions:

1. An approach where both epistemic uncertainty and aleatory uncertainty about component failure probability are modeled. In the proposed method, the epistemic uncertainty is modeled with a distribution whose parameters are also described by a

distribution.

2. A method that is founded in the hierarchical Bayesian model (HBM) to address current challenges of data scarcity, leverage potential availability of data in the future, and update the probabilistic evaluation of infrastructure performance.

3. A method to incorporate Bayesian updating of fragility formulations of damage to infrastructure components under earthquake and flood hazard. The updating of component fragility is integrated into the serviceability assessment of infrastructure systems.

2.2 Background

Much of the work done on the performance analysis of WDS focuses on earthquake and man-made hazards and evaluates component-level and system-level performance separately. For instance, at the system level, Wang and O'Rourke [57] simulate the response of the WDS under seismic hazard where system serviceability is defined as the ratio of satisfied demand after the earthquake to total demand before the earthquake. Another study uses a connectivity-based model to evaluate the seismic vulnerability of WDS based on Monte Carlo simulation and a shortest-path algorithm [56]. At the component level, Shuang et al. [58] propose a model to evaluate the nodal vulnerability of WDS under cascading failures due to intentional attacks. Fewer research studies focus on the mathematical model for the performance analysis of WDS under flood hazard, even though floods have caused more damage and fatalities than any other natural hazard in recent years, making it a particularly destructive and dangerous hazard [59, 60]. Current performance analyses of WDS under floods are based on HAZUS-MH [1][61, 62]. However, damage to buried pipes is typically not considered in HAZUS-MH since it is assumed that submergence of these pipes cannot

---

[1]HAZUS-MH is a nationally applicable, standardized methodology provided by the Federal Emergency Management Agency (FEMA) to estimate potential loss from multiple types of hazards (earthquakes, floods, hurricanes, and tsunamis), and visualize the impact of such hazards [5].

occur [5]. This assumption can be invalid as reports on the long-term recovery efforts in the wake of Hurricane Katrina have shown damaged pipe infrastructure in New Orleans for drinking water and wastewater transport. This finding contradicts the typical risk analysis assumption that floods do not damage buried critical infrastructure [63].

A challenge in evaluating the post-disruption performance of WDSs is the need for a large amount of data to estimate parameters such as the component damage state and build a reasonably accurate model with statistical methods [64, 16, 65]. However, due to the nature of extreme events and the limited access to buried infrastructure components, data collection during and after a disruptive event is challenging. As a result, there's uncertainty about the component vulnerability, network topology, and system serviceability after a disruption. Uncertainty is categorized as either aleatory or epistemic. While aleatory uncertainty is generally viewed as irreducible, epistemic uncertainty is associated with model selection and parameter assumptions, and can be reduced to improve the accuracy of the model [15, 18]. Failing to account for the epistemic uncertainty in the performance evaluation of critical infrastructure systems could lead to biased estimates of the system state, thereby causing prolonged failures and increased social and economic losses. As such, it is critical to characterize the uncertainty in the assessment of the serviceability of disrupted infrastructure systems. In order to address uncertainty, probabilistic models founded in Bayesian approaches have been developed and widely applied to assess system performance, reliability, risk, and resilience of critical infrastructure systems [66, 67, 68, 69, 70].

Bayesian methods allow for uncertainty quantification and integrate direct and indirect sources of information to improve the estimation accuracy [71, 16]. In the case of WDSs, existing studies that have employed that Bayesian approach are primarily focused on predicting WDS component failure to support vulnerability assessment of the WDS [72, 73, 74, 55, 54]. Specifically, Dong et al. [54] apply Bayesian networks to evaluate the vulnerability of flood control infrastructure using node failure probability and failure cascade susceptibility as the metrics. Additionally, modeling uncertainty has been consid-

ered by Kabir et al. [55] in a study that integrates Bayesian model averaging and Bayesian proportional hazard model to predict the failure of water mains. Bayesian updating is a promising approach under epistemic uncertainty because it updates model parameters and outputs as relevant data become available. Specifically, hierarchical Bayesian models (HBMs) are able to integrate information from various sources and adapt to the structure of data, thereby reducing the variance of estimates [75] and addressing concerns of defining a prior distribution with a very limited amount of data [16]. HBMs have been widely applied to risk and reliability analysis of different engineering systems [76, 16, 77, 78, 79, 65]. In particular, Yan and Haimes [16] demonstrated the ability of HBM to improve estimate accuracy through "strength borrowing" by pooling data from similar and related systems and applying HBM to handle the lack of data in risk-based system analysis.

While prior work has contributed to advances in system level performance and component level probabilistic modeling, few studies have explicitly considered the epistemic uncertainty around component fragility and its integration into the serviceability assessment of the entire WDS. This research proposes a Bayesian approach integrated with simulation and network algorithms to (i) consider the uncertainty in component failure prediction, and (ii) integrate the component metrics in a system-level performance assessment using Monte Carlo simulation and shortest-path algorithm. The approach is founded in HBM to enable updating of system-level performance given updated parameters of component-level metrics.

2.3 Methodology

This section presents the methodological components of the framework as they apply to a WDS starting with an overview of failure rate estimation methods and followed by a detailed description of component failure modeling and network performance evaluation.

2.3.1 Overview of Failure Rate Estimation Methods

2.3.1.1 Repair Ratio

Repair ratio that is calculated as the ratio between the number of repairs to a unit length of pipe is commonly used as the indicator for the impact of seismic hazard on pipelines [48]. In this chapter, the seismic fragility model from the American Lifelines Alliance (ALA) [80] is adapted to a more general form shown in Eq. (2.1) in order to model multiple types of hazards. For earthquake hazard, historical data collected from records of repairs to service lines and water mains after recorded seismic intensities are generally used to fit a regression model to estimate RR (per 1,000 feet of pipe length) as a function of hazard intensity I_c at the site of a component. For seismic intensity, PGV (in/s) is typically used. Aleatory uncertainty about PGV is represented by a random residual term described further in Section 2.3.3, so PGV is assumed to be a random variable. Of the collected data, pipe material, pipe joint type, pipe diameter, and soil condition also impacted the measured RR values, so a fragility curve modification factor K was defined to scale the final RR result from the regression model (Eq. (2.1)). HAZUS adopts the linear model from ALA to estimate the damage to water distribution systems given a seismic intensity from a specific earthquake scenario [4].

$$RR = aKI_c \tag{2.1}$$

In Eq. (2.1), a is the regression parameter and I_c is the intensity of the event. The factor K was originally developed for seismic events and is assumed here to be applicable to flood hazard as well.

To reflect flood intensity, a random variable h_f, measured in units of feet, represents the standing water depth in a certain location and will reflect I_c in Eq. (2.1). Historically, HAZUS provides floodplain extent and utilizes flood depth to calculate the damage to above-ground components such as pumping stations; however, recent studies document

HAZUS missing as much as 75% of flooded areas when validating the predicted versus actual flood extent of historical events [81]. The accuracy of this methodology depends on the accuracy of the flood depth estimates, which is largely determined by the quality of data that can be collected in the future. Ideally, repair records after major flood events should be collected and analyzed similarly to the existing methodology for earthquakes. After repair reports are collected and analyzed, new values for *a* and *K* are estimated through regression analysis to calculate the expected *RR* from flood events. With no database of records in existence yet, the initial values for *a* and *K* are adjusted from earthquake scenarios for floods.

The *RR* calculated from Eq. (2.1) is one of the main parameters used to estimate the probability of pipe failure. However, equation (2.1) is fit from a relatively small sample of data points on earthquake events. Additionally, there is no data available on pipe failures after flood events. The cost of collecting and analyzing such data is high due to sensitivity and security as well as the rare occurrence of disasters. Therefore, point values are often used for *a* and *K* even though the model itself has a poor fit, especially for large observed *RR* values [82]. To address this limitation, epistemic uncertainty is introduced to the variables *a* and *K* and a Bayesian updating approach is used to estimate RR.

2.3.1.2   Bayesian Updating and Hierarchical Bayesian Model

A Bayesian approach follows Bayes' theorem to update the estimate of the parameters of interest. Given a prior distribution of parameters (variables), $p(\boldsymbol{\theta})$, and a data likelihood function, $p(\boldsymbol{D}|\boldsymbol{\theta})$, the posterior density, $p(\boldsymbol{\theta}|\boldsymbol{D})$, is given by Eq. (2.2).

$$p(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{D}|\boldsymbol{\theta})}{p(\boldsymbol{D})} \propto p(\boldsymbol{\theta})p(\boldsymbol{D}|\boldsymbol{\theta}) \tag{2.2}$$

A Bayesian model can be dynamically updated where the prior distribution is iteratively replaced with the informative posterior distribution obtained using existing data when new data become available.

An extension of this approach is the hierarchical Bayesian model (HBM) which considers the uncertainty in the parameters of the prior distribution by adding another layer of Bayesian inference. Instead of setting a point value to the distribution parameters of $\boldsymbol{\theta}$, a hyperprior, represented by ϕ, is used to quantify the uncertainty around $\boldsymbol{\theta}$. According to the marginal posterior of ϕ shown in Eq. (2.4), ϕ is now partially determined by the dataset rather than being an assumed value [16].

$$p(\boldsymbol{\theta}, \phi | \boldsymbol{D}) \propto \underbrace{p(\boldsymbol{D}|\boldsymbol{\theta})}_{\text{likelihood}} \underbrace{p(\boldsymbol{\theta}|\phi)}_{\text{prior}} \underbrace{P(\phi)}_{\text{hyperprior}} \tag{2.3}$$

$$p(\phi|\boldsymbol{D}) = \int p(\boldsymbol{\theta}, \phi | \boldsymbol{D}) d\boldsymbol{\theta} \propto \int p(\boldsymbol{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\phi) P(\phi) d\boldsymbol{\theta} \tag{2.4}$$

Since Eq. (2.3) usually does not have a closed-form solution, simulation techniques are typically leveraged to generate samples of the posterior distributions of interest. Markov Chain Monte Carlo (MCMC) methods such as the Metropolis-Hastings algorithm, Gibbs sampling, and Hamilton (or Hybrid) Monte Carlo have been developed to perform sampling efficiently [83, 84]. MCMC algorithms coupled with the emergence of freely available software and inexpensive computing power have led to tremendous advances in Bayesian inference [85]. In this chapter, Stan, a probabilistic programming language that implements full Bayesian statistical inference [86], is employed to implement the proposed HBM. For more details about HBMs and sampling methods for HBMs, readers are referred to Ref. [75] and references therein.

In order to estimate RR under uncertainty and limited data, the variables a and K are assumed to be lognormal random variables with uncertain means, λ_a and λ_K, that follow a normal distribution. The standard deviations are assumed to be fixed such that the normal distribution is weakly-informative [87]. If no prior information is available, the hyperpriors are typically assumed to be weakly-informative distributions to build a more robust model [88]. Modeling details of the hazard intensity I_c will be covered in Section 2.3.2. The HBM for estimating *RR*, visualized in Fig. 2.1 with the corresponding mathematical

model given by Eq. (2.5) to Eq. (2.10), can be continuously updated with new pipe repair records to ensure the most accurate parameters are utilized to dynamically inform resource allocation during a disaster as well as for long-term planning.

$$RR \sim \mathcal{N}\left(\mu_{RR}, \sigma_{RR}\right) \tag{2.5}$$

$$\mu_{RR} = aKI_c \tag{2.6}$$

$$\ln a \sim \mathcal{N}\left(\lambda_a, \zeta_a\right),\ \ln K \sim \mathcal{N}\left(\lambda_K, \zeta_K\right) \tag{2.7}$$

$$\ln I_c = \ln \bar{I}_c + \ln \varepsilon_I,\ \ln \varepsilon_I \sim \mathcal{N}\left(\lambda_\varepsilon, \zeta_\varepsilon\right) \tag{2.8}$$

$$\lambda_a \sim \mathcal{N}\left(\mu_a, \sigma_a\right),\ \lambda_K \sim \mathcal{N}\left(\mu_K, \sigma_K\right) \tag{2.9}$$

$$\sigma_{RR} \sim \mathcal{N}\left(\mu_\sigma, \sigma_\sigma\right),\ \zeta_a \sim \mathcal{N}\left(\mu_\zeta, \sigma_\zeta\right),\ \zeta_K \sim \mathcal{N}\left(\mu_\zeta, \sigma_\zeta\right) \tag{2.10}$$
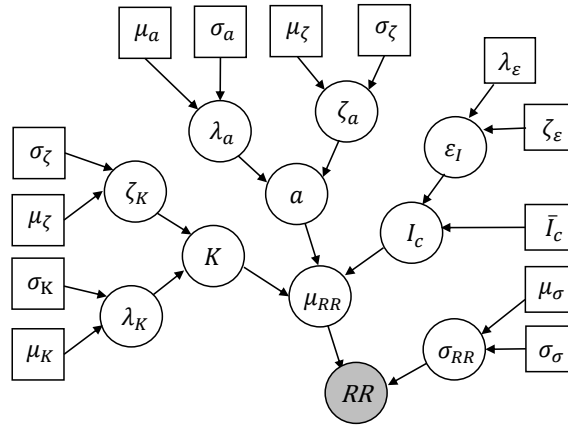
Figure 2.1: HBM for estimating *RR*. Circles denote random variables. Squares denote constants. Shaded nodes denote observations while unshaded ones denote hidden variables/constants.

2.3.2 Hazard Intensity

The variable I_c from Eq. (2.1) must be determined for each component of the network to calculate individual failure probabilities. For earthquake scenarios, the hazard intensity I_c

23

can be measured in terms of peak ground velocity *PGV* (cm/s) or peak ground acceleration *PGA* (cm/s^2). *PGA* and *PGV* are estimated from their respective attenuation equation that models seismic intensity at a site from an earthquake of magnitude M_w and distance from the epicenter *R* (km). The attenuation equations for the median *PGV* and *PGA* adopted in previous research are shown in Eqs. (2.11) and (2.12) [56, 82].

$$\log_{10}(\overline{PGA}) = 3.79 + 0.298 \times (M_w - 6) - 0.0536 \times$$
$$(M_w - 6)^2 - \log_{10}(R) - 0.00135 \times R \tag{2.11}$$

$$\log_{10}(\overline{PGV}) = 2.01 + 0.422 \times (M_w - 6) - 0.0373 \times$$
$$(M_w - 6)^2 - \log_{10}(R) \tag{2.12}$$

The median seismic intensity *PGV* from Eq. (2.12) is used to describe the damage to pipes (link components) while the median *PGA* from Eq. (2.11) is used to describe damage to facilities (node components). The standard deviations of the residuals associated with Eq. (2.12) and Eq. (2.11) to capture the aleatory uncertainty around seismic intensities are typically assumed to be a lognormal distribution with a median value equal to 1.0 and a standard deviation of 60% [56]. The intensity of a seismic event varies across the system and is usually modeled as a homogeneous two-dimensional stochastic field with a residual term [89]. The estimation equation for the logarithmic residual of the seismic intensity ($\overline{PGA}$ or $\overline{PGV}$) at the site of a component *c*, denoted by $\ln I_c$, is given by Eq. (2.13) where $\bar{I}_c$ represents the mean value of I_c . The autocorrelation function of logarithmic residuals for the seismic intensities at the site of component c_i and c_j is given by Eq. (2.14) where R_c is the distance from the site of component *c* to the epicenter, the correlation distance *b* is the strength of the spatial correlation and is typically assumed to be 30 km [89].

$$\ln(\varepsilon_I) = \ln\left(\frac{I_c}{\bar{I}_c}\right) \tag{2.13}$$

$$\rho(\ln I_{c_i}, \ln I_{c_j}) = \exp\left(-\frac{||R_{c_i} - R_{c_j}||}{b}\right) \tag{2.14}$$

The flood intensity is measured by flood depth (feet) estimated by the flood depth based on the digital elevation model (DEM). The elevation of a given location is subtracted from the maximum water surface elevation along the cross-section of a flood basin. However, due to the slope of the floodplain and the complexity in flow paths, estimating flood depth using cross-section maximum flood water elevation is challenging [90]. A more accurate method to generate spatially-explicit floodwater depth is through numerical simulation given data on the hydrological characteristics and riverine morphology [90]. Hydrology-based flood map applications include FEMA Flood Maps [91] and the National Weather Service Inundation Maps [92] from the National Oceanic and Atmospheric Administration (NOAA). In the study of this chapter, we adopt a simulation-based methodology using HAZUS. The HAZUS Flood Model uses characteristics such as frequency, discharge, and ground elevation to model the spatial variation in flood depth and velocity. It is important to note that HAZUS can underestimate the flood depths by considering the precipitation in the study region without accounting for cascading effects from upstream rivers. To offset the underestimation, a longer return period can be considered in HAZUS so that the simulated flood depths are approximate to observations at the gauges within the study area. The HAZUS Flood Model is subject to several sources of uncertainty, such as the variation in channel and floodplain elevation in the Digital Elevation Model (DEMs) and the variation in floodplain extent and depth in the River Hydraulic Model [93]. To characterize the resulting uncertainty around the flood depths, a 50% standard deviation is assumed for estimates of flood depths. To improve the accuracy of these estimates, the average of the estimates from two simulations can be used. If the flood depth estimates at some components of the WDS are missing in one simulation due to the removal of problematic reaches, the estimates from the second simulation can be used. If the missing values from the two simulations overlap (The data for these reaches are not available in HAZUS), the mean of the estimates derived from elevation-based interpolation can be used.

2.3.3 Component Failure Modeling

Each WDS is represented by a network consisting of nodes and links. Given the hazard intensity and RR estimates, a probability of failure is derived for each node and link in the WDS.

2.3.3.1 Node Failure

The physical damage to facilities in a water distribution network such as elevated storage tanks or pumping stations is described with fragility curves and are used in HAZUS [4]. A total of five damage states are defined, including none (ds_1), minor (ds_2), moderate (ds_3), extensive (ds_4), and complete (ds_5). As an example, the fragility curve for above-ground steel tank entering different damage states is shown in Fig. 2.2. In the study of this chapter, damage state ds_5 is adopted.
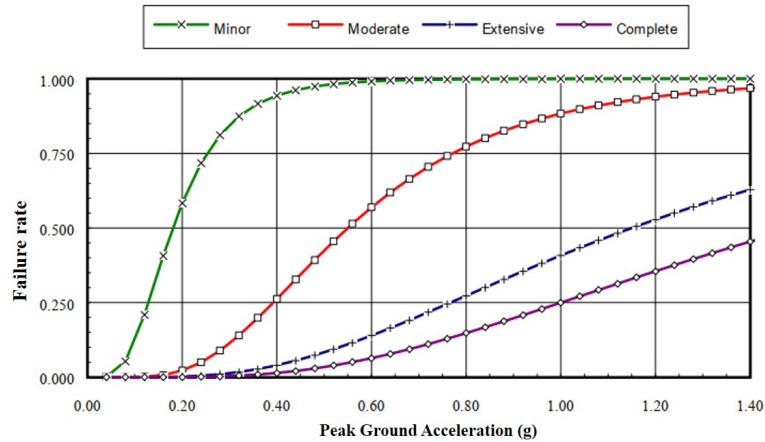
Figure 2.2: Fragility curve for above-ground steel tank under earthquake [4]. Failure rate represents the probability of a component reaching a specific damage state given *PGA*.

The fragility curve is described with a log-normal distribution as a function of *PGA* given in Eq. (2.15) where P_v describes the failure probability of each node, λ_{PGA} represents the logarithmic mean of *PGA* that is measured in the gravitational acceleration *g*, ζ_{PGA} describes the standard deviation of $\ln(PGA)$, and $\phi(\cdot)$ is the standard normal cumulative

distribution function.

$$P_v\left(PGA\right) = \phi\left(\frac{\ln\left(PGA\right) - \lambda_{PGA}}{\zeta_{PGA}}\right) \tag{2.15}$$

The intensity of a flood event is described by the flood depth. Flood fragility curves given the flood depth are typically derived from historical data, damage survey data, or expert opinions [94]. Since fragility curves of water facilities are not available in the literature, historical data of the pumping stations (medium/large, above ground) and storage tanks (all, above ground) from the HAZUS Flood Model technical manual [5] are leveraged to fit the fragility curves presented in Fig. 2.3. The two curves are used to evaluate the failure probability of water facilities in the case study of Shelby County, TN. Note that the failure probability of the elevated storage tanks is always 0 when the flood depth is under 10 feet. For other types of pumping stations or storage tanks, the fragility curves can be derived in the same manner.
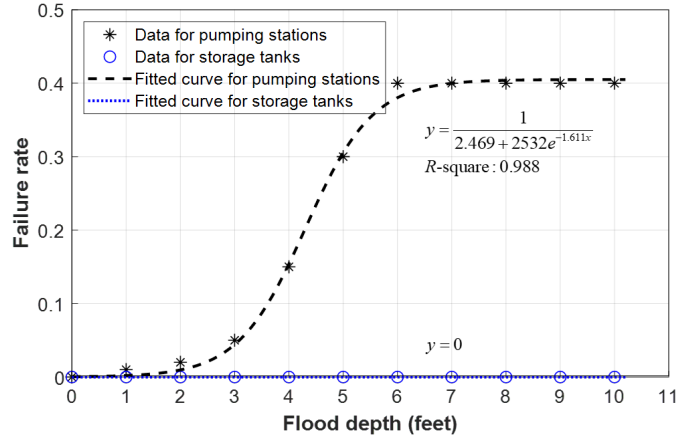
Figure 2.3: Fragility curve for water facilities under flood

2.3.4 Link Failure

The damage to buried pipes is discussed more generally in Section 2.3.1.1. As the hazard intensity, PGV or flood depth, varies along a pipe, the pipe is equally divided into m segments. First, the repair ratio RR_i for a segment L_i is assessed. Then the number

27

of pipe breaks within this segment, n, is modeled using a Poisson distribution as shown in Eq. (2.16), where ΔL_i is the segment length [80]. The failure probability of a pipe, P_l, complementary to the probability that no breaks occur in all of the pipe segments, is provided in Eq. (2.17).

$$P_{L_i}[n = k] = \exp\left(-RR_i \times \Delta L_i\right) \times \frac{(RR_i \times \Delta L_i)^k}{k} \qquad (2.16)$$

$$P_l = 1 - \exp\left(-\sum_{i=1}^{m} RR_i \times \Delta L_i\right) \qquad (2.17)$$

2.3.5 Network Performance

In order to evaluate the serviceability of a water distribution network under a disruption scenario, a double-loop MCS is employed to generate multiple scenarios. The approach is illustrated in Fig. 2.4 and follows 9 steps.

1. The natural hazard scenario is generated. For earthquake hazards, the epicenter and magnitude of the earthquake are defined. For flood scenario, the severity of the event based on the return period, such as a 100-year or 200-year flood event, is determined.

2. The intensity of the natural hazard, i.e. *PGA* or *PGV* for earthquake or the standing water depth for flood at the location of vulnerable network components, is estimated respectively.

3. A random vector of correlated hazard intensities for each component is generated.

4. The probability of failure for each component of the network is calculated. HBM is used to obtain the component failure probability. For simplicity, the components are assumed to be either fully functional or inoperable.

5. The status of a component under the hazard scenario is determined by comparing its failure probability to a random number $r \sim U(0,1)$. If the failure probability is

28
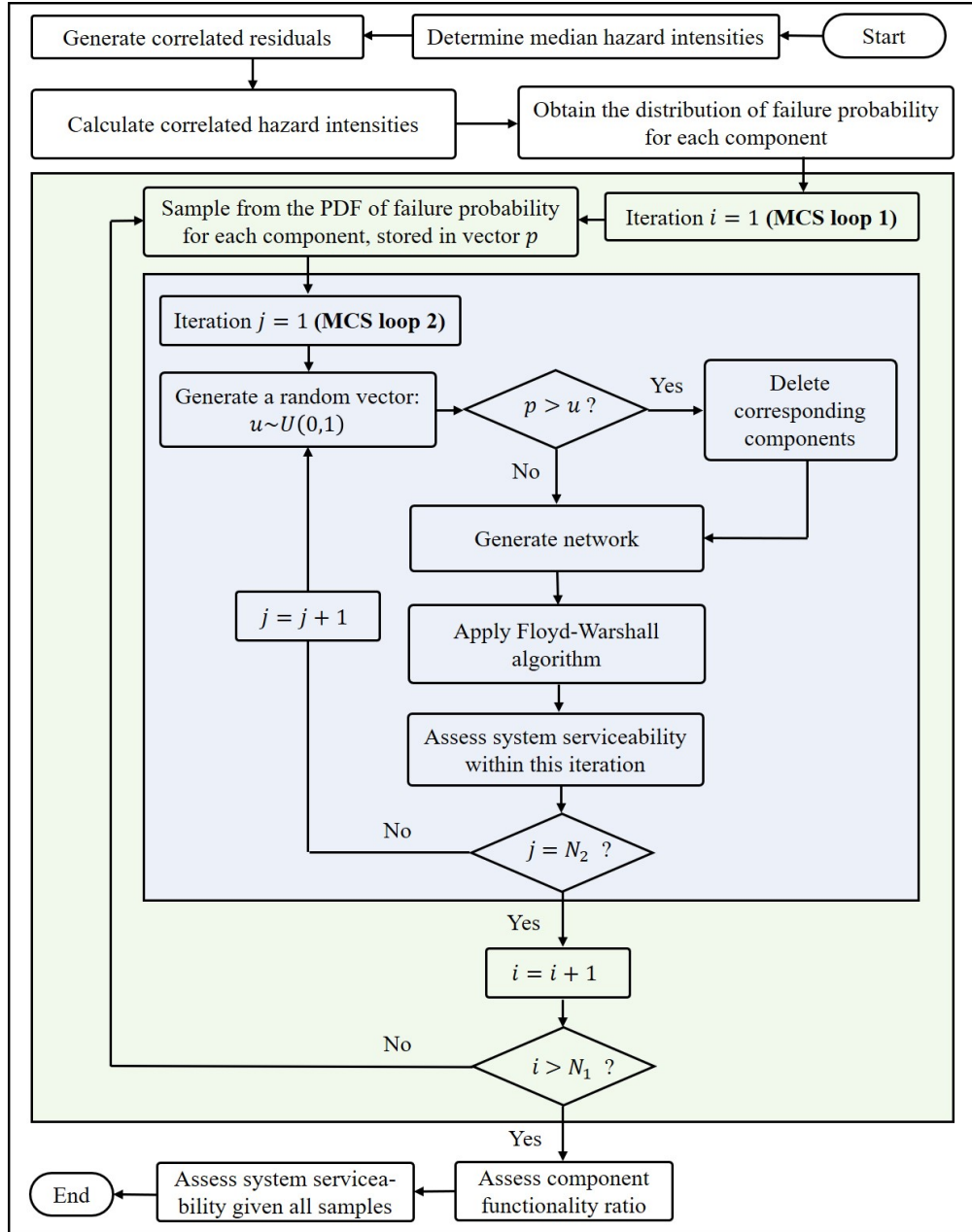
Figure 2.4: Flowchart of the proposed approach

greater than r, then the component is considered damaged and is removed from the network and a subgraph of the original network is generated.

6. The functionality of components and network serviceability are evaluated using the Floyd-Warshall algorithm [95, 96], which finds the shortest path between all pairs of nodes simultaneously. In this algorithm, the distance between disconnected nodes is

equal to infinity. If at least one path exists from a demand node to a source node, the demand node is considered to be functional.

7. Steps 3-6 are performed for N_2 times to develop a probability distribution of component functionality and the network serviceability.

8. Steps 3-7 are repeated for N_1 times to capture the uncertainty in system serviceability due to the epistemic uncertainty associated with a and K.

9. Once the functionality ratio of the demand nodes is obtained, network serviceability, s, is evaluated using Eq. (2.18) where F_{n_i} is a binary variable describing the functionality of each demand node i (1 indicates functionality and 0 otherwise); N_d is the total number of demand nodes.

$$s = \frac{\sum_{i=1}^{N_d} F_{n_i}}{N_d}, \ s \in [0,1] \tag{2.18}$$

When epistemic uncertainty is introduced in step 7, a distribution of the mean serviceability of the network is generated rather than a single point estimate of the average serviceability described in Eq. (2.18). This distribution allows for an understanding of the impact of uncertain model parameters on the outcome of disruption scenarios.

2.4   Case Study

2.4.1   Test Network

To illustrate the proposed approach, a case study of a real-world WDS in Shelby County, Tennessee is presented in this section. The WDS (Fig. 2.5) serves approximately one million people. The network consists of 6 elevated storage tanks, 9 pumping stations, 34 relay nodes, and 71 buried water pipes. The relay nodes are the demand nodes while the storage tanks are the supply nodes. Since the relay nodes constitute the branch points where the water pipes intersect, the damage to these nodes is not considered in the study of this

chapter. The pipes installed before 1975 were made from cast iron with molten lead joints (before 1959) or mechanical joints (1959-1975) while the pipes installed after 1975 were made from ductile iron pipes with slip joints. The pipe diameter ranges from 6 inches to 48 inches [97].
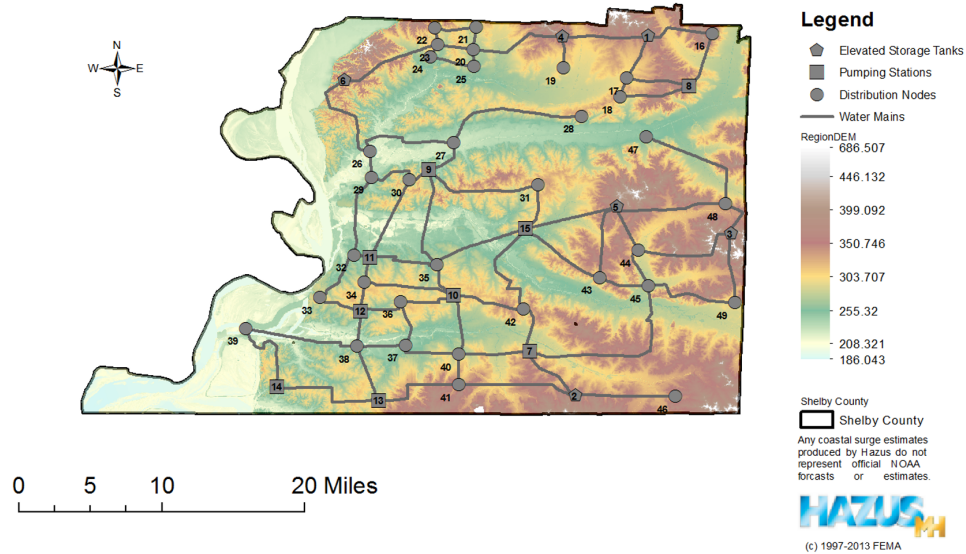
Figure 2.5: WDS of Shelby County [5]

2.4.2   Natural Hazard Scenario

2.4.2.1   Earthquake

The study area is earthquake-prone because the New Madrid Seismic Zone is centered northwest to Shelby County. The maximum probable earthquake with an exceedance probability of 2% in 50 years centers at 35.3 N and 90.3 W [98]. The distance of the WDS components to the epicenter for the maximum probable earthquake ranges from 20 km to 65 km, with a mean value of approximately 40 km. Once the earthquake scenario is defined, the median *PGA* at a point is solely dependent on the distance to the epicenter. Therefore, the median *PGA* contours in Shelby County with the default data in HAZUS given the earthquake scenario above show a rippling shape (Fig. 2.6), decreasing gradually

from the highest value of approximately 0.6 in the northwest corner to the lowest value of around 0.3 in the southeast corner.

2.4.2.2 Flood Hazard Scenario

Shelby County borders the Mississippi River on the west, and several smaller rivers run across the county. As such, Shelby County is particularly vulnerable to flooding. The blue area in Fig. 2.7 shows the 100-year HAZUS-derived flood zone. In order to match the 100-year flood based on the records of river crests in Memphis [99], a 1000-year flood is simulated in HAZUS.

2.4.3 Results

2.4.3.1 Serviceability Assessment

The serviceability assessment is performed following the approach presented in Section 2.3.5. The correlation distance b is set to 30 km for earthquakes while correlation is not considered in calculating the flood depth at different sites of network components. For
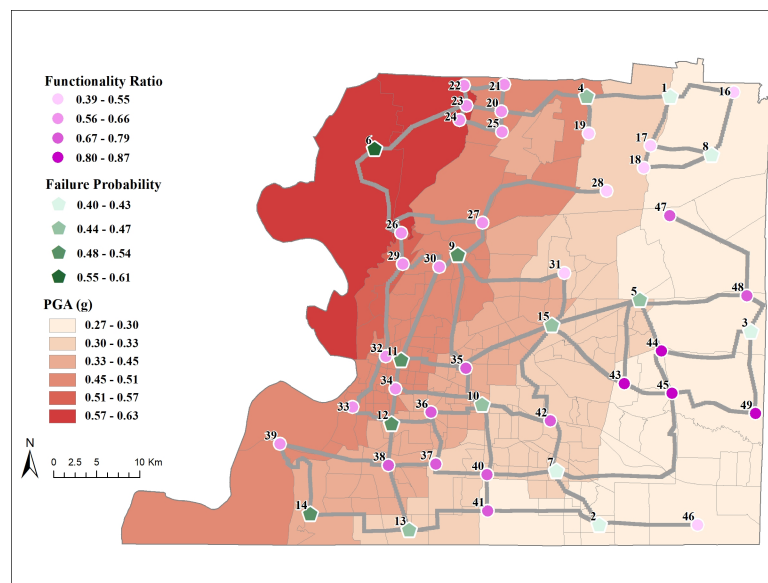
Figure 2.6: Functionality of relay node and failure probability of storage tanks and pumping station vs. median *PGA* in Shelby County

32

damage caused by earthquakes and assuming a damage state ds_5, $\lambda_{PGA} = \ln(1.5)$ is used for both water storage tanks and pumping stations, and $\zeta_{PGA} = 0.6$ is used for water storage tanks and $\zeta_{PGA} = 0.8$ for pumping stations [4]. In calculating the repair rate of water pipes under seismic activities, $\mu_K = \ln(0.5)$ and $\mu_a = \ln(0.00187)$ are assumed based on the point estimate $K = 0.5$ and $a = 0.00187$ from ALA. For the damage caused by flood, $\mu_K = \ln(0.5)$ is also assumed and the coefficient μ_a is assumed to be $\ln(0.001)$ so that the magnitude of the failure probability of water pipes under flood is close to that under earthquake. Given no prior information, standard deviations, ζ_a, ζ_K, σ_a, and σ_K are set to be 5 to make the respective distribution weakly informative [100]. These assumptions are crude due to the lack of data on *RR* under flood hazard. Once new data about *RR* become available, the assumptions can be modified to make the distributions more informative.
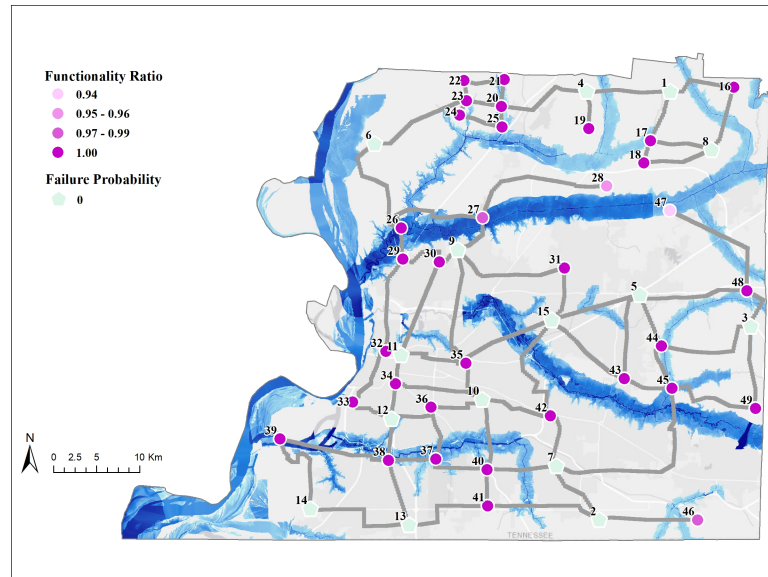
Figure 2.7: Failure probability of water distribution facility and functionality ratio of relay nodes under flood (The blue area is the flood zone)

In the simulation, N_1 and N_2 are set to 1000, therefore a total of 1.0×10^6 simulation runs are performed to obtain the mean functionality ratio of each relay node under earthquake and flood hazard. As the component fragility is distinct under earthquake and flood events, the distribution of functionality ratio under different hazards are disparate. Figure 2.8 indicates that the functionality ratios of relay nodes are much higher under flood
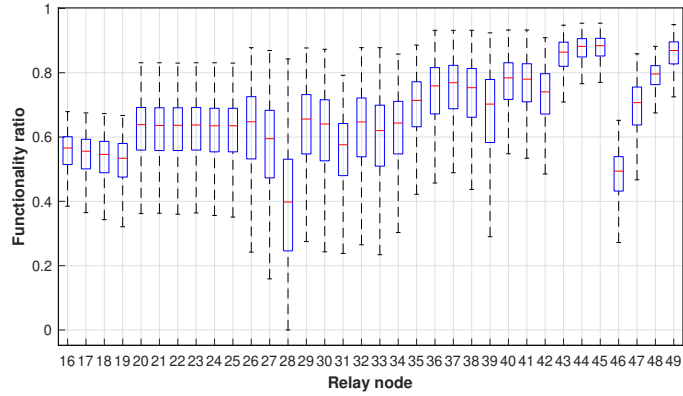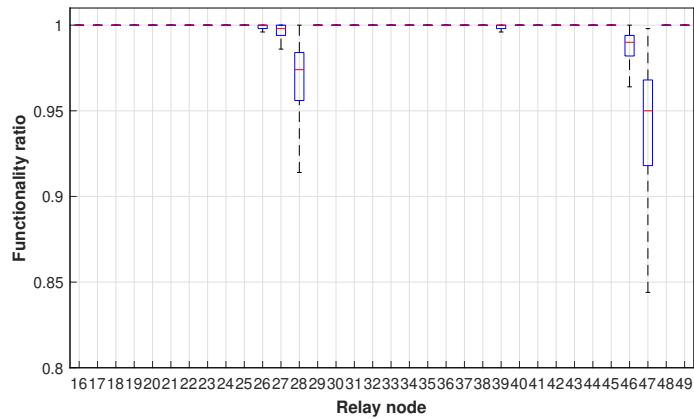
Figure 2.8: Component functionality ratio under (a) earthquake and (b) flood

than under earthquake. Under seismic hazard, the functionality ratios range between 0.4 and 0.9 while under flood, most of the functionality ratios are close to 1.0. The reason is that the components are more vulnerable to seismic hazard than to flooding because during a flood event, components located at high altitude are not inundated (Fig. 2.7) while under an earthquake event, all the components are subject to the impact of seismic waves (Fig. 2.6). In particular, node 28 has a low functionality ratio under both hazards because it is only accessible from one water pipe and it is far from the water storage tanks. In comparison, node 46 and node 47 have higher functionality ratios because they are close to storage tanks. However, node 47 has the lowest functionality ratio under flood hazard because it is closely located to the river and it is accessible through only one water pipe. As

such, node functionality is dependent on the fragility of nodes and on the topology of the system. These observations indicate that system topology is more important in determining the functionality ratio under seismic hazard while component fragility is the dominant factor under flood hazard.

The respective probability density function (PDF) of the mean serviceability ratio under earthquake and under flood is presented in Fig. 2.9. This matches the results of functionality ratio as the mean serviceability ratio is much higher under flood than under earthquake. Compared to the PDF under earthquake, the PDF under flood is narrow. This small variance is due to the fact that only a small portion of components fail under flood, leading to a small number of different subgraphs constructed by removing inoperable nodes and links in the MCS.
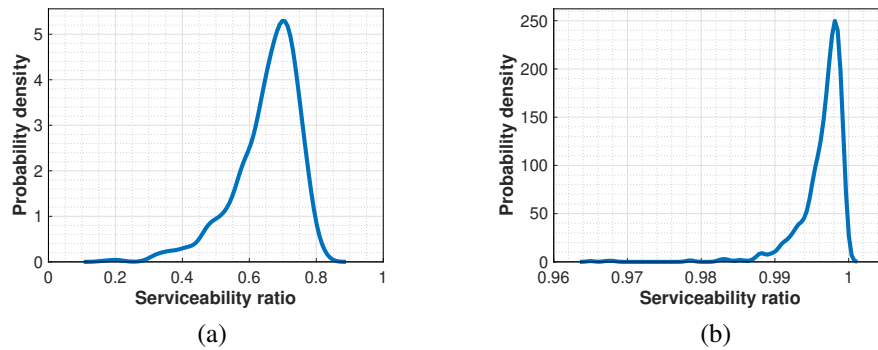
Figure 2.9: PDF of the mean serviceability ratio under earthquake (a) and flood (b)

2.4.3.2 Node Importance Analysis

Node importance analysis is a crucial step in infrastructure management as critical nodes functionality can significantly impact the serviceability of the entire network [101], thereby guiding decision-making on resource allocation and prioritization of repair activities when multiple nodes are disrupted [69, 102]. To obtain the component importance ranking, a method based on Kullback-Leibler (KL) divergence [103] is adopted. KL divergence is often used as a measure of dissimilarity between two probability distributions

in information theory [104]. In this chapter, KL divergence is applied to calculate component importance based on the probability distributions of network serviceability obtained by changing the component functionality. A sample of the distribution of network serviceability is generated through the proposed simulation approach and KL divergence from one discrete probability distribution, P, to another discrete probability distribution Q, is used as shown in Eq. (2.19), where T is the number of possible values for the two discrete distributions; $P(i)$ and $Q(i)$ are the i-th probability mass in their respective distribution.

$$D_{\mathrm{KL}}(P\|Q) = -\sum_{i=1}^{T} P(i) \log \frac{Q(i)}{P(i)} \qquad (2.19)$$

The importance of a component is equal to the KL divergence between the probability distributions of the network serviceability before and after deleting that component, denoted as S_0 and S_c, respectively. The importance of component i, denoted by w_{c_i}, can be computed according to Eq. (2.20).

$$w_{c_i} = D_{\mathrm{KL}}(S_0\|S_{c_i}) \qquad (2.20)$$

When the importance of nodes that represent storage tanks and pumping stations is evaluated using MCS, each node is removed individually to reveal its contribution to the decrease in the network serviceability. The steps for ranking components based on KL divergence are summarized in Algorithm 1.

Algorithm 1: Rank components based on KL divergence

1: Run the double-loop MCS to generate the initial distribution of system serviceability

   S_0 given all components with the respective failure probability

2: **for** $i = 1$ to N **do**

3:     Remove node i

4:     Run the double-loop MCS to obtain the serviceability distribution S_{c_i}

5:     Calculate component importance with Eq. (2.20)

6: **end for**

7: Sort w_{c_1} to w_{c_N} in descending order
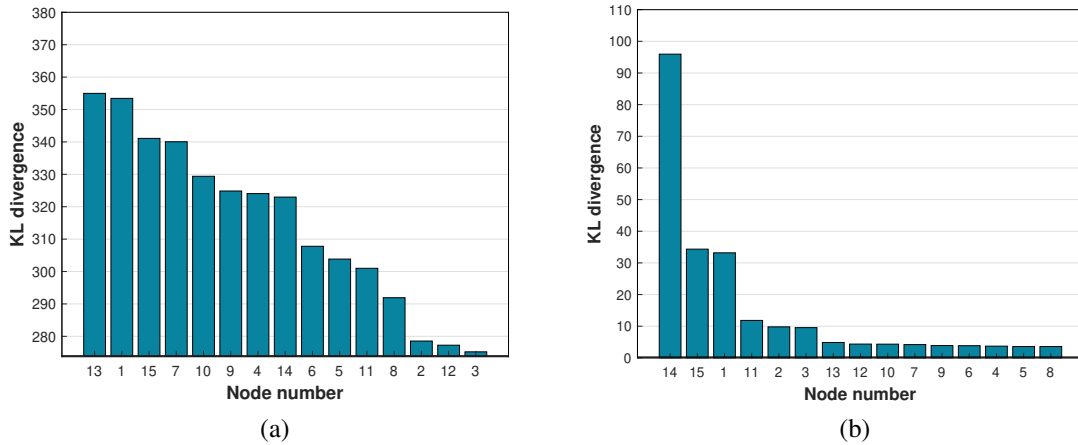
Figure 2.10: Node importance under earthquake (a) and flood (b)

The results in Fig. 2.10 show that the ranking of node importance under earthquake is different from that under flood. Under earthquake, nodes 13 and 1 have the highest influence on the overall system serviceability while node 2, 12, and 3 have the least influence on the serviceability. Under flood hazard, node 14 is by far the most important to the system serviceability. Node 4-10, 12 and 13 have a trivial impact on the serviceability.

2.5   Summary

This chapter presents a Bayesian updating framework, including a HBM for network component failure rate, to assess the impact of natural hazards on infrastructure networks. The uncertainty about component fragility due to data scarcity and randomness in hazard intensities is measured using Bayesian updating models. Uncertainties about components are propagated to the system level using MCS to assess the network serviceability under the impact of different types of hazards. The proposed framework is illustrated with a case study of a real-world WDS in Shelby County, Tennessee. The results indicate that the WDS performs differently under different hazard scenarios. The network is more vulnerable to earthquake hazard than flood hazard. Further, component functionality is driven by network topology under earthquake hazard whereas under flood hazard component vulnerability is more important. The difference is also reflected in the ranking of node importance under the two hazard scenarios. The ranking of components can be used by utility managers and emergency responders to inform the allocation of resources in disaster preparedness and response.

Chapter 3

Hierarchical Bayesian Kernel Method for Quantifying Resilience Under Small Data

3.1 Introduction

In the aftermath of a natural disaster, a loss in the functionality of infrastructure systems can severely disrupt the infrastructure systems and the daily life of customers they served. Effective management of recovery requires accurate estimation of resilience metrics which allows for informed allocation of resources before, during, and after a disaster, in order to reduce the loss caused by hazards that hit the infrastructure systems. As such, quantifying the recoverability of such services is therefore highly critical to the recovery of infrastructure systems.

As is stated in Chapter 1, the capacity to recover from the occurrence of a disruptive incident is an important component of system resilience. In order to accurately quantify the recoverability of infrastructure systems, it is important to identify metrics of evaluation and their corresponding definitions. Definitions of full recovery vary in the literature, however, they can be summarized into three categories, (a) returning to "pre¬-disaster" conditions, (b) attaining what would have occurred 'without' the disaster, or (c) reaching a new stable state that may be different from either of these [105]. In this chapter, full recovery, depending on the type of disasters, type of indicators, and characteristics of communities, is defined as either returning to pre-disaster conditions or reaching a new state that can be better or worse than the pre-disaster conditions.

To date, although various metrics have been developed to measure the post-disaster recovery for different types of hazards, little research has been conducted to apply them to quantifying post-disaster recovery. Liu et al. [106] develop an approach to assess the recovery of infrastructure systems from earthquakes considering four dimensions (population, economic, building, and infrastructure) by extending the concepts of the resilience

triangle, yet the method presented in their work is qualitative as the recovery is expressed using a score. The work presented in this chapter constitutes a first step in the quantification of resilience of infrastructure systems in which the recovery metrics are measured as a function of time or number of people impacted.

A common issue in estimating the metrics for the recovery of infrastructure systems is the lack of recovery data from historical events. Traditional statistical learning tools rely on large datasets in order to provide robust estimates with a high level of accuracy [16]. As such, scarce data situations result in challenges in measuring and predicting resilience metrics within an acceptable level of accuracy using existing statistical models. In this chapter, the rate of recovery in the aftermath of a disaster is used as the metric for resilience. This metric has been defined and used as a key indicator of measuring the recovery of infrastructure systems in prior studies on disaster [106]. This work develops a new statistical learning method, the hierarchical Bayesian kernel model (HBKM), which integrates the Bayesian property of improving predictive accuracy as data are dynamically obtained, the kernel function that can make nonlinear data more manageable, and the hierarchical property of borrowing information from different sources in scarce and diverse data situations which are common in disaster scenarios. This method is applied to quantifying the recovery rate from power outages and is compared to other statistical methods for validation.

The objective of this chapter is to improve the estimation of the recovery state of infrastructure systems under scarce data situations by developing a novel statistical model. Cross-validation techniques are employed to validate the new model against classical models. In particular, the log-likelihood and Root Mean Squared Error (RMSE) are used as the metrics for predictive accuracy. An illustrative case study is presented in which the models are applied to power recovery data collected from five of the most severe storms that ever hit Shelby County in Tennessee since 2007.

3.2   Preliminaries

3.2.1   Hierarchical Bayesian Methods

Hierarchical methods are comprised of multiple parameters that are representative of the structure of the problem or the data whereby a joint probability model for these parameters is used to reflect their dependence [75]. More specifically, in order to estimate the parameters of interest, hierarchical Bayesian (HB) methods employ a multi-level model in which Bayes' theorem is utilized to integrate the sub-models under each level with the data to account for the uncertainty of the parameters of interest [107]. The result of this integration is the posterior distribution, which is a two-level HB model that can be represented graphically with Fig. 3.1 or mathematically with Eq. (3.1)

$$p\left(\theta_j|\phi,y_j\right) = \frac{p\left(y_j|\phi,\theta_j\right)p\left(\theta_j,\phi\right)}{p\left(\phi,y_j\right)}, \tag{3.1}$$

where ϕ represents the hyper parameter. y_j represents one group of data that follows a distribution with parameter θ_j. Each θ_j in the prior is described by a hyperprior distribution with a set of hyperparameters, denoted as ϕ. Note that the distributions of different groups of data y_j depend on ϕ only through θ_j. In the non-hierarchical model, ϕ is assumed to be deterministic and known, and it is represented by one value. In the hierarchical model, however, the uncertainty of ϕ is considered and accounted for using a probability distribution. This multi-level structure reduces the subjectivity of making assumptions about the priors by assigning a distribution to the prior parameter and updating it with observed data. The hyperpriors are typically specified by using expert elicitation, incorporating existing information, or using a proper non-informative distribution. Further details on specifying prior distributions can be found in Ref. [87].

HB models offer several advantages over non-hierarchical models. First, HB models have enough parameters to fit the data well, thereby avoiding problems of overfitting and
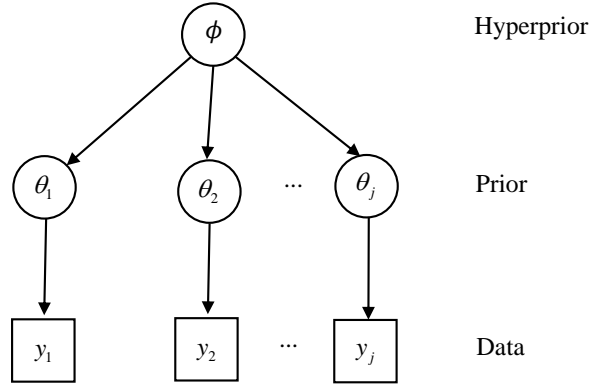
Figure 3.1: Graphical Representation of a HB model.

improving prediction accuracy [75]. Additionally, HB models are flexible in combining information across different groups of data. As such, HB models are widely employed in probabilistic risk analysis. This feature of HB models is referred to as "strength borrowing" and can be demonstrated using analytical derivations of the posterior calculations of the parameters [16]. Suppose θ_j cannot be obtained within an acceptable level of confidence due to data scarcity. HB models address this problem by borrowing strength from data collected from a similar subset of the hierarchical structure. All θ_j's are governed by the same underlying distribution which will be updated with the entire dataset $\boldsymbol{y}$. Consequently, the posterior distribution of θ_j does not depend solely on the direct data point y_j, but also the commonly updated hyperprior, as can be seen from Eq. (3.2) [16].

$$
\begin{aligned}
p\left(\theta_j|\phi,\boldsymbol{y}\right) &= \frac{p\left(\boldsymbol{y}|\phi,\theta_j\right)p\left(\theta_j,\phi\right)}{p\left(\phi,\boldsymbol{y}\right)} \\
&\propto p\left(y_j|\phi,\theta_j\right)p\left(\theta_j,\phi\right) \\
&= p\left(y_j|\theta_j\right)p\left(\theta_j|\phi\right)p\left(\phi\right)
\end{aligned}
\tag{3.2}
$$

While the HB method is powerful, the solution techniques are often computationally intensive, especially when the hierarchical models are too complicated to generate a closed form of posterior distributions. As such, MCMC methods help in these situations to utilize HB models by replacing the analytical solution with repetitive calculations that computers

42

can perform over a large number of iterations [107]. Recent advancement in computational tools and efficiency of Monte Carlo algorithms resulted in HB models becoming more applicable and popular [108]. In this chapter, Rstan, a package in R for Bayesian modeling and inference [100], is employed to infer the HB models that are considered in the case study.

3.2.2 Bayesian Kernel Models

Bayesian kernel methods are a class of models in which kernel functions are integrated with Bayesian analysis, producing probabilistic rather than deterministic estimates of the parameter of interest, and thus accounting for uncertainty in the underlying data [109]. The approach was first developed in machine learning with the Gaussian kernel model which assumes a Gaussian prior for the response variable [110]. The covariance matrix is used as the kernel matrix.

Extensions of the Gaussian kernel model to non-Gaussian cases make the Bayesian kernel model applicable to a wider range of problems. Montesano and Lopes [111] develop the beta kernel model that uses a beta distribution as the conjugate prior distribution to predict the probability of a robot successfully grasping an object. MacKenzie et al. [109] highlight two issues associated with Gaussian kernel models in binary classification, imbalanced datasets, and longitudinal datasets. Specifically, the complexity of Gaussian kernel models limits their ability to update parameters quickly as new data become available. They improve and generalize the beta Bayesian kernel model to deal with heavily imbalanced datasets by adding weights to the kernel matrix in the computation of the posterior. It is found that the beta kernel model consistently outperforms the Relevance Vector Machine (RVM) and LASVM (an incremental learning version of Support Vector Machine) if 50 or fewer data points are available, and the model frequently performs better than the RVM and LASVM even if more data are available [109, 112].

Baroud et al. [113] develop the Poisson Bayesian kernel model that integrates prior

information from experts' domain knowledge and historical data to calculate the probability distribution of the rate of occurrence of a particular event. The model assumes that the rate of occurrence λ, which has a Poisson likelihood function, follows a Gamma(α, β) prior distribution. Because Gamma distribution is a conjugate prior for the Poisson distribution, the posterior distribution of λ is still a Poisson distribution. The posterior of λ is estimated based on the kernel matrix that measures the similarities of the attributes between the test data (new data) and the training data (historical data). The kernel function is represented by K_j which is the $n_j^{\text{test}} \times n_j^{\text{train}}$ kernel matrix for the j-th group of data, $\boldsymbol{y_j}$ is a $n_j^{\text{train}} \times 1$ vector containing the output data associated with n_j^{train} observations of $\boldsymbol{X}$, and V_j is an $n_j^{\text{train}} \times 1$ vector of ones. The posterior distribution of λ is then Gamma (α_j', β_j') with the parameters of the posterior computed using Eqs. (3.3) and (3.4).

$$\alpha_j' = \boldsymbol{K_j} \boldsymbol{y_j} + \alpha_j \tag{3.3}$$

$$\beta_j' = \boldsymbol{K_j} \boldsymbol{V_j} + \beta_j \tag{3.4}$$

With the kernel matrix, the posterior of λ is estimated based on the similarities of the attributes between the new data point and the training data (historical data). The parameters α_j and β_j in the prior distribution of λ are deterministic and are determined based on expert elicitation or empirically using the method of moments. A graphical representation of the Bayesian kernel model is given in Fig. 3.2. In this graphical representation (and all graphical representations shown thereafter in this research), the directed arrows represent direct influence between the nodes, the square nodes represent deterministic values, and the round nodes represent probabilistic variables. For this particular Poisson Bayesian kernel model displayed in Fig. 3.2, the dataset $\boldsymbol{y}$ is divided into J groups with n_j data points in each of them. Within the j-th group, y_{ij} is governed by λ_{ij} whose posterior distribution is determined by parameters α_j and β_j and the kernel matrix $\boldsymbol{K}$. The kernel matrix is calculated using the value of predictors and the tuned kernel parameter σ_j of each group.
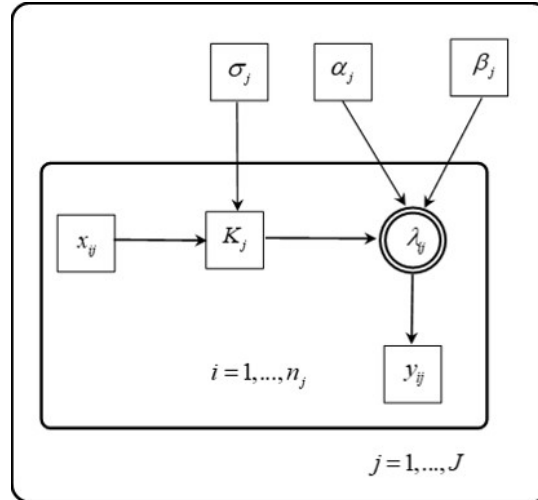
Figure 3.2: Graphical representation of the Bayesian kernel model

Later, Baroud [114] conducted a comparative analysis of the Poisson Bayesian kernel model, Poisson and Negative Binomial generalized linear model using the several goodness-of-fit and prediction accuracy metrics. The results show that Poisson Bayesian kernel model appears to be a good model for prediction purposes when the data set is small with a small number of predictors.
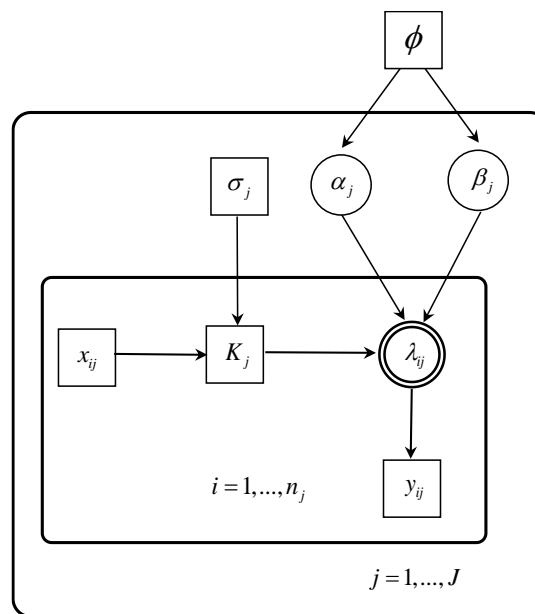
Figure 3.3: Graphical representation of the hierarchical Bayesian kernel model

3.3　Methodology

3.3.1　Hierarchical Bayesian Kernel Model

HBKM integrates the HB model with the kernel function in order to allow for strength borrowing and to incorporate attribute information [17]. Compared to the Bayesian Kernel Model (BKM), the HBKM improves the accuracy of the estimation of α_j and β_j in the prior distribution of λ by describing them using an extra layer of Bayesian updating. As a result, the values of α_j and β_j are probabilistic and are estimated using the corresponding full dataset. The graphical representation of HBKM is shown in Fig. 3.3 where α_j and β_j are shown in a round node to indicate that they are random parameters described by the same hyperprior ϕ, whereas in Fig. 3.2, α_j and β_j are shown in a square node to indicate that they are scalars. Note that the dataset does not necessarily need to be grouped in the BKM because α_j and β_j are a scalar.

The HB model used to obtain the updated hyperparameters in HBKM is given in Eq. (3.5) where ϕ represents the common hyperprior following a half-Cauchy distribution.

$$
\begin{aligned}
&y_{ij} \sim \text{Pois}(\lambda_{ij}),\ i=1,2,...,n_j,\ j=1,2,...,J \\
&\lambda_{ij} \sim \text{Gamma}(\alpha_j,\ \beta_j) \\
&\alpha_j \sim \text{Cauchy}(0,\ 5),\ \beta_j \sim \text{Cauchy}(0,\ 5)\ (\alpha_j > 0,\ \beta_j > 0)
\end{aligned}
\tag{3.5}
$$

The radial basis function is chosen to be the kernel function given by Eq. (3.6)

$$
K\left(x_{ij}^{\text{test}}, x_{ij}^{\text{train}}\right) = e^{-\frac{\left\|x_{ij}^{\text{test}} - x_{ij}^{\text{train}}\right\|^2}{2\sigma_j^2}}
\tag{3.6}
$$

where x_{ij}^{train} and x_{ij}^{test} represent the i-th data points in the training set and test set of the j-th group, respectively. This function is widely used as it generates a kernel matrix with (i) all elements within zero and one, which facilitates the interpretation of a similarity measure between the attributes of data points, (ii) full rank [110], and (iii) only one free parameter

to be tuned to the optimal value, hence reducing computation time [114]. The parameter σ_j in the kernel function is tuned by maximizing the log-likelihood function given by

$$L\left(\hat{\lambda}_j\right) = \sum_{i=1}^{n_j}\left[y_i\ln\left(\hat{\lambda}_j\right) - \hat{\lambda}_j - \ln\left(y_i!\right)\right] \tag{3.7}$$

where the final term can be ignored in the tuning process since it does not depend on the parameter of interest.

In the hierarchical version of the Poisson Bayesian kernel model, Eqs. (3.3) and (3.4) become Eqs. (3.8) and (3.9) respectively

$$\alpha'_j = \boldsymbol{K_j}y_j + \alpha_j^* \tag{3.8}$$

$$\beta'_j = \boldsymbol{K_j}\boldsymbol{V_j} + \beta_j^* \tag{3.9}$$

where α_j^* and β_j^* now represent the mean values of the updated distributions of α_j and β_j, respectively.

3.4   Case Study

3.4.1   Data Description

The proposed model is applied to a case study in which the resilience of the power system is defined as the number of customers without power. The goal is to model the stochastic recovery of an infrastructure system as a function of time after a storm. The models are applied to assessing the recovery from power outages in Shelby County, Tennessee. The entire population of Shelby County is served by the main utility company serving approximately 500,000 customers is Memphis Light, Gas, and Water (MLGW), the largest three-service municipal utility in the the U.SUSA. Shelby County, which includes Memphis, the second largest city in Tennessee, is vulnerable to hazards including earthquakes, flooding, and storms. Assessing the recoverability of infrastructure systems
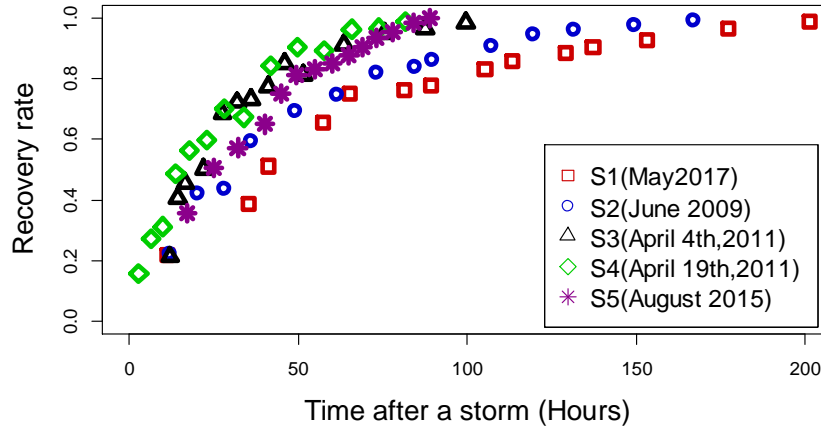
Figure 3.4: data on power recovery rate for the five most severe storms

can assist local agencies and the utility in protecting customers against adverse impacts of disasters by improving preparedness for future events. In this case study, the recoverability of power after storms will be explored to assist the utility company in recovery management.

Power outage data are collected from MLGW news releases which are posted on the company's website and updated intermittently during and after a storm. The dataset includes the number of people without power given the time after each of the 15 documented storms that hit Shelby from 2007 to 2017. In the present study, data from the five most intense storms, referred to as S1, S2, S3, S4, and S5, are chosen to be analyzed using the different models.

Figure.3.4 illustrates the recovery rate of power in Shelby Count after each of the five storms considered in this case study. As noted earlier, the proposed method is applicable to any definition of recovery rate. In this specific example, the original data on the number of people without power at the given time is translated into the number of people without power per thousand, then the model yields the predicted number of people without power per thousand at the desired time, and the recovery rate is calculated as $1 - C/1000$ where C represents the number of customers without power obtained from the model. Note that after S1 and S2, the two strongest storms, the power system took almost double the time

to achieve full recovery. Table 3.1 summarizes the initial number of people without power after each storm as well as the number of recorded data points on the recovery process.

Table 3.1: Initial number of people without power and size of data set in each storm

Storm	Initial number of people without power	Number of data points
S1 (May 2017)	188000	14
S2 (June 2009)	133000	9
S3 (April 4th,2011)	63500	6
S4 (April 19th, 2011)	64000	6
S5 (August 2015)	67842	6

Since RStan does not support missing values in the data matrix, non-linear interpolation is used to impute the missing data points and ensure that all data subsets have the same structure and size. The non-linear interpolation is based on the exponential or logarithmic model fitted for each subset of data representing one storm. Gaussian noise is added to each interpolated data point in order to minimize the sensitivity of the model performance to the interpolation approach and to have the imputed data conform closely to reality where the observations will usually fluctuate around a true model. The mean of the Gaussian noise is 0 and the standard deviation is 10% of the fitted value.

3.4.2 Comparative Analysis

In order to investigate the performance of HBKM, a comparative analysis is conducted where the proposed method model is compared with two other statistical models, GLM and HBRM. The assessment is carried using cross-validation, where a subset of the data is held out while the models are fitted to the rest of the data and then tested on the held out test set. Poisson GLM is a classical approach to deal with count data such as the occurrence rate of an incident. Each data point of y is fit with a Poisson distribution, and the value of y given new attribute information is estimated using the mean rate of occurrence λ. The mean rate λ is thus the parameter of interest and the log of λ is expressed as a linear function of predictors, x. In the case of one predictor, the coefficients θ_0 and θ_1 in the model are

assumed deterministic and are calculated using Least Squared Method. Note that the data points are fully independent and need not to be grouped in GLM. For the convenience of comparison to the previous two models, the data are still considered as grouped in the graphical representation of the GLM. The corresponding equation is given by

$$y_{ij} \sim \text{Pois}\left(\lambda_{ij}\right), i = 1, 2, ..., n_j, j = 1, 2, ..., J \tag{3.10}$$

$$\log\left(\lambda_{ij}\right) = \theta_0 + \theta_1 x_{ij} \tag{3.11}$$

The other model used in the comparison is HBRM. This model, expressed as in Eq. (3.12) where n_j is the number of data points in j-th group of data set, also assumes that the log of mean rate is related to predictors, x, through a linear relationship. However, the coefficients in the log-linear function are assumed to be probabilistic and follow a non-informative normal distribution. Half-Cauchy$(0,5)$, a weakly-informative distribution recommended by Gelman et al. [87], is used to model the hyperparameters. The coefficients θ_j in the log-linear model for the rates in different groups of data set vary according to each group. The coefficients are indirectly correlated through the common hyperprior, and the intercept θ_0 remains unchanged in all the groups. MCMC is used employed to compute the posterior distribution of the parameters.

$$y_{ij} \sim \text{Pois}\left(\lambda_{ij}\right), i = 1, 2, ..., I, j = 1, 2, ..., J \tag{3.12a}$$

$$\log\left(\lambda_{ij}\right) = \theta_0 + \theta_j x_{ij} \tag{3.12b}$$

$$\theta_0 \sim N\left(\mu_0, \tau_0\right), \theta_j \sim N\left(\mu_1, \tau_1\right) \tag{3.12c}$$

$$\mu_0 \sim N(0,5), \mu_1 \sim N(0,5) \tag{3.12d}$$

$$\tau_0 \sim \text{Cauchy}(0,5), \tau_1 \sim \text{Cauchy}(0,5) \quad (\tau_0 > 0, \tau_1 > 0) \tag{3.12e}$$

3.4.3 Performance Assessment

In this case study, the cross-validation technique is used adopted to evaluate the out-of-sample predictive error. More specifically, 35% of the data are used as the training set, 35% as the tuning set to optimize the parameter in the kernel function, and 30% is used as the test set, the. The samples are randomly selected over 100 iterations. In performing MCMC for the HB models, four chains are deployed with 3000 iterations in each chain where 30% of the iterations are for warmup and 70% are for sampling. All the models are also compared to having no model (null model), which simply uses the mean value of the in-sample data set, i.e. the training and tuning set in each group, as the out-of-sample predictions for the test set in each group. Four model performance metrics are computed: (i) log-likelihood, (ii) RMSE, (iii) relative frequency of the highest log-likelihood, (iv) the relative frequency of the lowest RMSE across all iterations. The mean, Probability Density Function (PDF), and Cumulative Density Function (CDF) of the RMSE and log-likelihood are used utilized to compare the performance of each model. In evaluating the models based on the CDF of the log-likelihood and RMSE, second-order stochastic dominance (SSD) is applied to compare the distribution of the values of these metrics. SSD provides a comprehensive technique to compare CDFs. For two random variables X and Y with corresponding CDF F_X and F_Y, if $F_X \leq F_Y$ (s) holds for any real number $s \in \mathbb{R}$, then X has first-order stochastic dominance over Y, denoted as $Y \preceq X$. However, if Eq.(3.13) holds, then X has second-order stochastic dominance over Y, denoted as $Y \preceq^2 X$.

$$\int_{-\infty}^{t} F_X(s)\,ds \leq \int_{-\infty}^{t} F_Y(s)\,ds \tag{3.13}$$

In the domain of risk-based decision making, a rational risk-averse decision-maker, who prefers the expectation of a random return $\mathbb{E}(X)$ to the random return X itself [115], will choose the alternative that leads to X if a higher value of X is preferred and $Y \preceq X$ or $Y \preceq^2 X$. To see why the first-order and second-order stochastic dominance imply risk-

51

aversion, interested readers are referred to Ref. [116]. In evaluating the predictive accuracy of the models considered in the case study, the model for which the CDF is dominated by the other models' CDFs of RMSE is the best since a smaller RMSE is preferred.

3.4.4   Results

3.4.4.1   Model Accuracy

The proposed model in this research will ultimately be utilized by disaster responders and decision-makers to understand how communities respond to and recover from disasters. As such, model should provide a high predictive accuracy whereby it is possible to forecast the recovery trajectory of an infrastructure system given information on the storm intensity depicted in this work by its length. This section investigates the predictive accuracy and computational time of the models using the power outage recovery data of Shelby County. Two evaluation metrics, the log-likelihood and Root Mean Squared Error (RMSE), are computed using the cross-validation approach outlined earlier. One iteration of the cross-validation uses 8400 samples to build the posterior distribution and calculates one value for each the RMSE and log-likelihood is calculated using the mean of the posterior distribution as a point estimate. The mean of 100 values of log-likelihood and the mean of 100 values of RMSE for the models considered are shown in Table 3.2 and the PDFs and CDFs of the two metrics for the models are provided in Fig. 3.5 and Fig. 3.6, respectively.

As can be observed from Table 3.2, for the Shelby County power recovery data, all three models exhibit much higher prediction accuracy than the null model with HBKM providing the best mean predictive accuracy. Specifically, HBKM has an average RMSE of 82.2, whereas HBRM and GLM have an average RMSE of 92.1 and 99.3, respectively. Having no model yields the largest mean RMSE.

While mean values of the evaluation metrics provide a general idea of the performance of the models, a more comprehensive and thorough way to assess these metrics across methods is to evaluate their distributions. From the CDFs of log-likelihood shown in Fig-
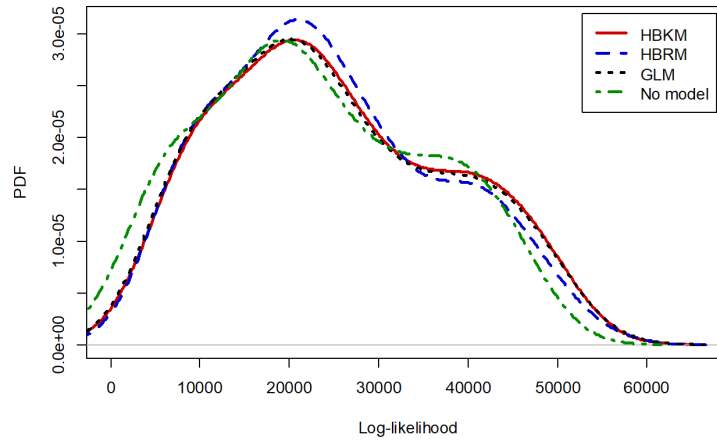
(a)
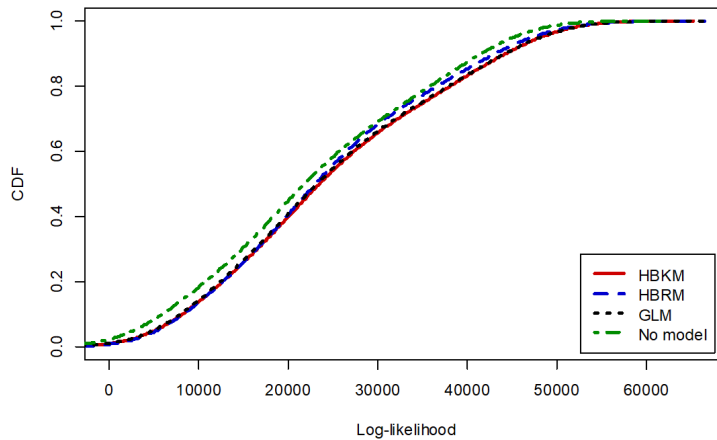
(b)

Figure 3.5: (a) PDF and (b) CDF of the log-likelihood for each model

ure 3.5, it can be seen that having a model outperforms the "no model" option because higher values of CDF are preferred and the CDFs of log-likelihood for the three models has first order dominance over the CDF of log-likelihood for "no model" (i.e., the CDF of log-likelihood for the "no model" is consistently larger than that of the other three models). However, differentiating between the three models' overall performance is not trivial and cannot be done visually. Fig. 8 displays the plots of the PDF and CDF for the RMSE of each model. GLM produces the lowest variance in the values of RMSE while HBRM has the largest variance. Similarly to the log-likelihood distributions, the CDFs of RMSE shows that having a model outperforms the "no model" because high values of RMSE are

53

(a)

(b)
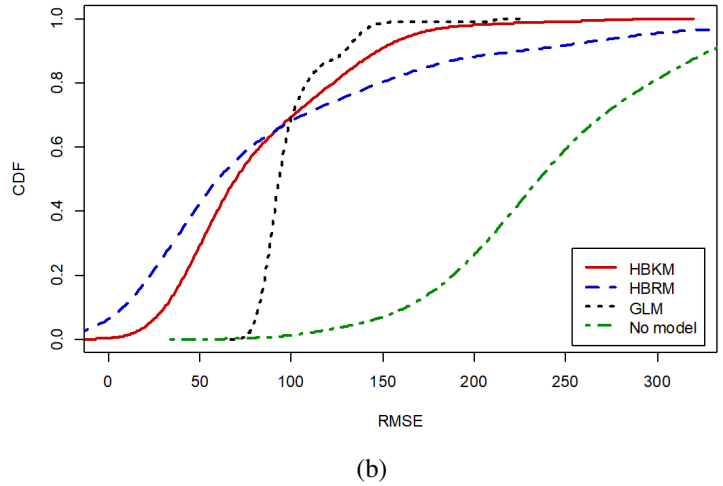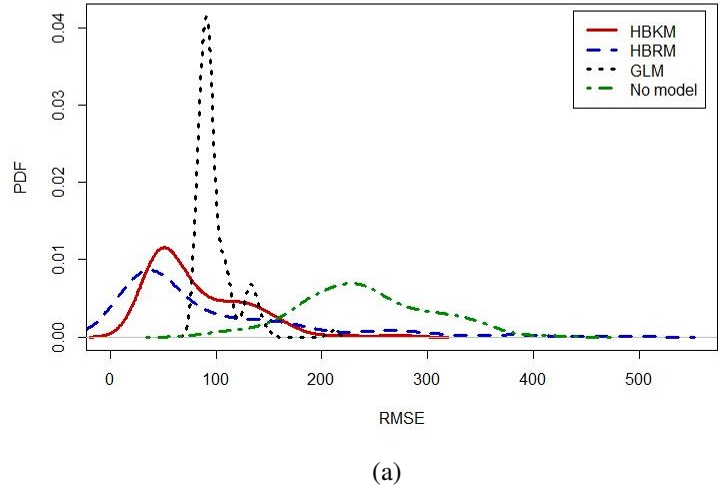
Figure 3.6: (a) PDF and (b) CDF of the RMSE for each model

not preferred and "no model" first-order stochastically dominates the three models (i.e., the CDF of "no model" is consistently smaller than that of the other models). However, among the three models, the dominance changes as a function of the value of RMSE whereby GLM dominates for values less than 100 while HBRM dominates otherwise.

In order to compare the overall performance of the models across all possible values of RMSE and log-likelihood, it is necessary to examine the integral of their CDF to determine the second order stochastic dominance. Figure 3.7 display the integral of the CDF of log-likelihood and RMSE. The integral of the CDF of the log-likelihood of HBKM is consistently smaller than that of HBRM and GLM, as a result, HBKM has second-order

Table 3.2: Comparison of mean prediction accuracy

Metric	HBKM	HBRM	GLM	No model
Log-likelihood	**25120**	24559.3	24900.5	23059.9
RMSE	**82.2**	92.1	99.3	240.8

dominance over the other models and is hence preferred. The integral of the CDF of RMSE for HBKM is larger than that of HBRM and GLM for values for RMSE that are larger than 200. As such, HBRM and GLM have second-order stochastic dominance over HBKM for large values of RMSE, indicating that while HBRM is more accurate for small values of error, HBKM obtains a better performance in terms of keeping the error within an acceptable range.

In addition to prediction accuracy, the computation time of the different models is examined. Specifically, the mean computation time for one iteration of each model averaged over 50 realizations is provided in Table. 3.3. It is clear that GLM and having "no model" provide the fastest computation results while the MCMC simulation for the hierarchical models takes a much larger amount of time. HBRM is the most computationally intensive method requiring slightly less than one minute on average for each iteration, while HBKM provides similar if not better predictive results for a significantly smaller amount of time. When such models are scaled up to larger systems and interconnected networks, the computation time will make a significant difference in the choice of the model.

Table 3.3: Comparison of average computational time

Model	Average computational time (s)	Standard deviation (s)
HBKM	34.6	0.5
HBRM	75.9	14.7
GLM	0.0033	0.0043
No model	0.0004	0.0031

Finally, the frequencies of achieving the best prediction out of the 100 iterations in the hold-out cross-validation are plotted in Fig. 3.8. According to this measure, HBRM
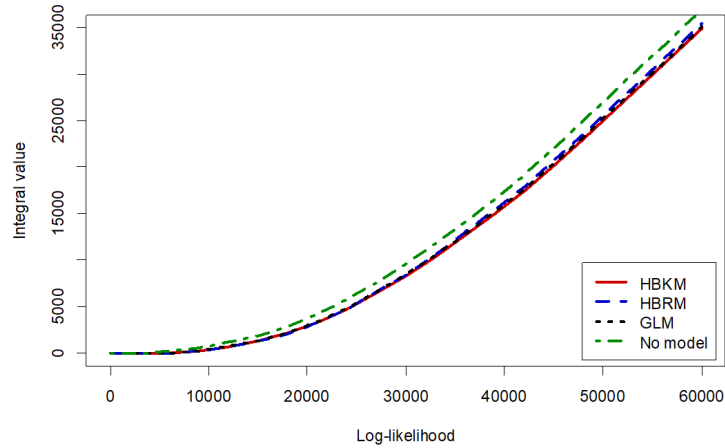
(a)
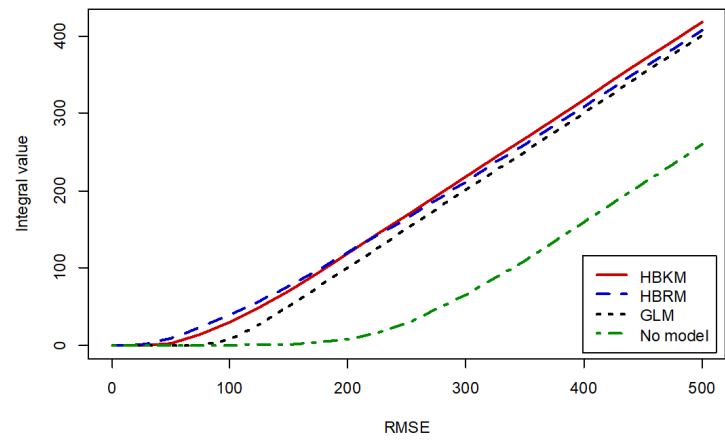
(b)

Figure 3.7: Integral of the CDF of (a) log-likelihood and (b) RMSE

provides the highest frequency of the best predictive accuracy metrics, i.e., smallest RMSE and largest log-likelihood, followed by HBKM, both outperforming GLM.

Overall, disaster responders' choice of model for carrying out the predictive analytics for the resilience of infrastructure systems will depend on a number of factors such as the model performance, the decision-maker's risk attitude, and the implementation of the model as a predictive and decision support tool. A decision-maker's risk attitude might impact the weight they place on prediction error as opposed to computation time or interpretability. While HBKM can be outperformed in a few instances by HBRM, HBKM requires less computational effort with little difference in the accuracy metrics.
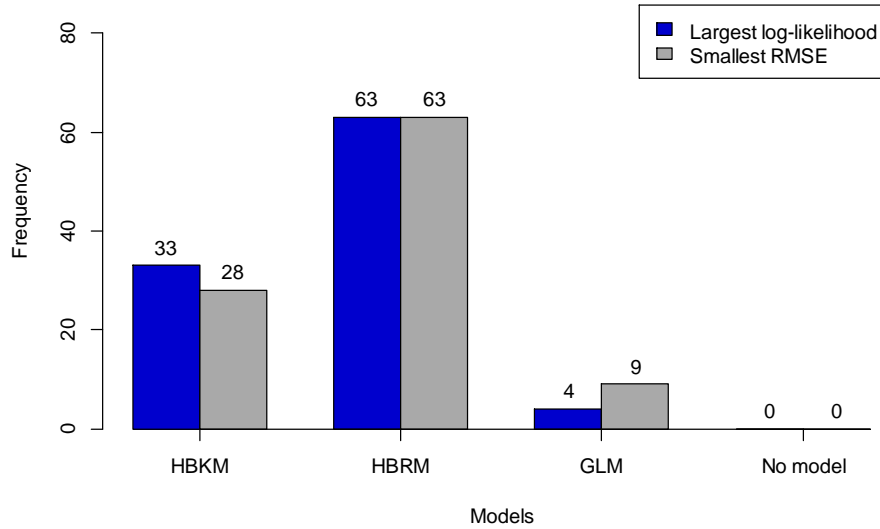
Figure 3.8: Comparison of the frequency of achieving the best performance

3.4.4.2 HBKM Model Outcome and Interpretation

In this subsection, possible model outcomes of the HBKM are discussed that would be helpful for a stakeholder when performing risk-based decision making post-disaster.

Figure 3.9 provides an example of the predictions of the recovery rate by the HBKM as a function time after a storm. The recovery rate for each storm is calculated by dividing the number of people who restore power at a given time by the initial number of people without power. At each point in time, a probability distribution of the possible number of customers regaining power is generated which provides the decision-maker with a range of values to consider for planning and recovery strategies. For illustration purposes, the predictions of storms S1 and S4 are presented. The predicted values follow a Poisson distribution with the predicted mean and observed values highlighted. From the distribution of predicted values, it can be observed that at the early stages of recovery after a storm, when the number of people without power is large, the predicted values have a relatively flat distribution, indicating a high uncertainty in the posterior distribution. As the infrastructure system is recovering and the number of people without power decreases, the predicted values get closer to the observed values and the posterior distributions become narrower, indicting
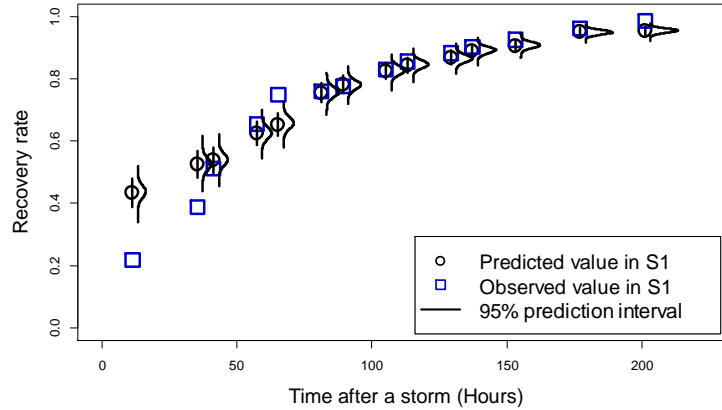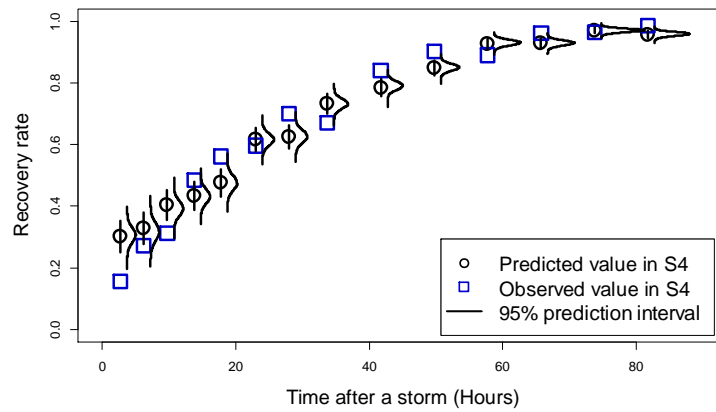
Figure 3.9: Comparison between observed values and predicted values by HBKM given time after a storm for (a) S1 and (b) S4 (The error bar represents the 95% confidence interval of a predicted value).

lower uncertainty in the predictions. Please note that the reduction of uncertainty around the predicted recovery rate in Poisson HBKM is due to the property of Poisson distribution since this distribution has a higher variance for a higher mean. However, investigating how the uncertainty of the predicted recovery rate changes over time will require applying other HBKMs whose likelihood distribution does not have a lower variance for a lower mean. The development of other HBKMs that employ non-conjugate priors is a direction of future work.

Using hold-one-out cross-validation, the observed and predicted values are plotted for all five storms in Fig. 3.10. Again, HBKM outperforms other models. Specifically, when
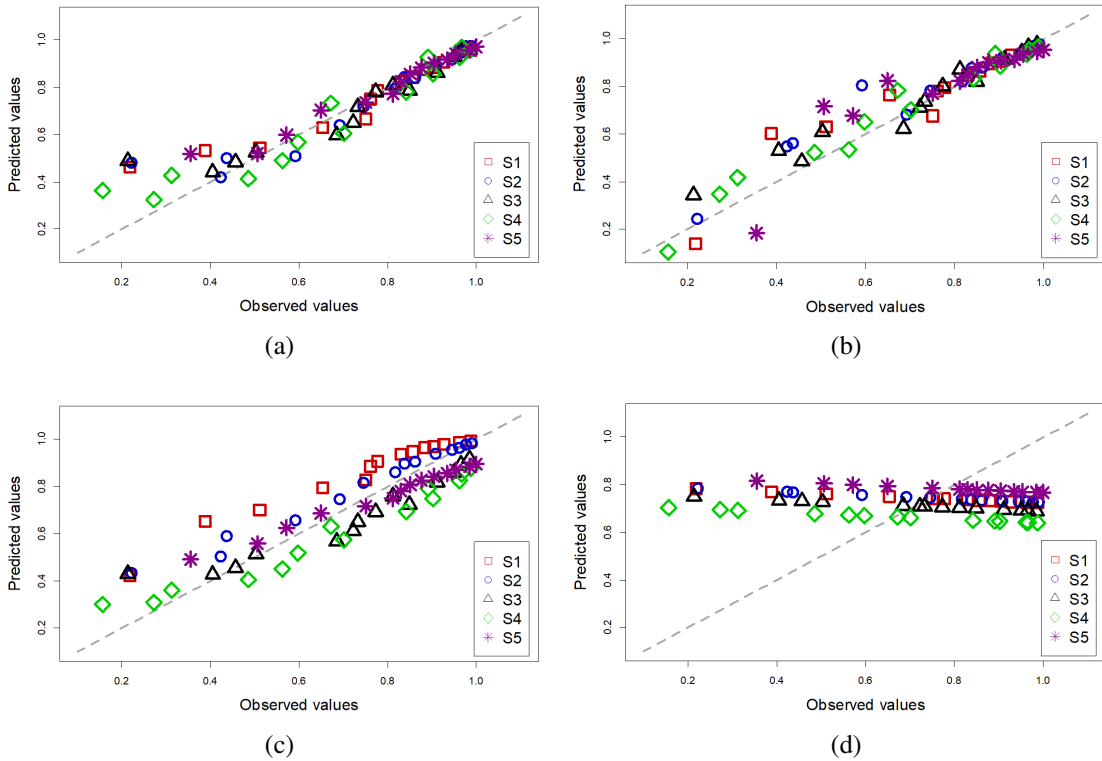
Figure 3.10: Comparison between predicted values and observed values for all storms. (a) HBKM (b) HBRM. (c) GLM. (d) No model.

compared to the results of other models, HBKM provides very accurate predictions when the recovery rate is larger than 40%. For smaller values of the rate of the recovery rate, the model has a tendency to overestimate the true value due to an underestimation of the number of people without power. However, recall from the previous section that HBKM outperforms all other models for larger values of RMSE, suggesting that in cases where there is a certain discrepancy between the observed and predicted value, the proposed model is still more accurate than the other statistical models considered in the analysis.

By providing an accurate estimate of the recovery rate of a certain infrastructure system, HBKM can be a powerful tool for utility companies and disaster responders in general. For utility companies, this model can help to (i) reduce the cost of investment in operations and maintenance by reducing call center volumes and managing vegetation along power lines, (ii) improve customer satisfaction by providing more accurate information to customers on

estimates of recovery, (iiiI) and more importantly, optimize the allocation of repair crews using accurate estimations of restoration time and the additional number of crews required through "mutual aid agreements" to ensure rapid restoration of electrical power to the affected areas [117]. Insufficient crews required due to an underestimation of the recovery rate will lead to prolonged recovery that increases the economic loss of the utility company. However, an overestimation of the recovery rate will also increase the cost on the part of the utility companies since they need to pay more for extra crews and logistics. The importance of accurate estimation of the number of crews required can be demonstrated numerically through an example (adapted from Ref. [118]). Suppose 30 components of an electric power system break down and 3000 residential customers lose power as a result. The utility managers need to decide on the number of additional repair crews required, apart from the three crews of their own. The cost of the three crews of their own is \$5,000 per day and the cost of additional crews and logistics combined per day is given by Eq. (3.14):

$$
f_c(n_a) = \begin{cases} \$10,000, & n_a = 1,2,3 \\ \$10,000 + \$5,000\,(n_a - 3), & 4 \leq n_a \leq 10 \end{cases} \tag{3.14}
$$

where n_a represents the number of additional crews and the maximum number of additional crews is 10. Next, for simplicity of demonstration, full restoration is assumed to be achieved in 10 days with the three original crews. That is, one crew can restore power to an average of 100 customers per day. However, the efficiency of additional crews is lower than the original crews and will decline with the increase in the number of additional crews. The efficiency of one additional crew f_e, defined by the number of customers that they can restore power to per day, is given by Eq. (3.15):

$$
f_e(n_a) = 100 - 5n_a, \quad 1 \leq n_a \leq 10 \tag{3.15}
$$

With $n_a + 3$ crews, the total time of full recovery is T days, which is given by Eq. (3.16):

$$T(n_a) = \left\lceil \frac{3,000}{3 \times 100 + n_a f_e(n_a)} \right\rceil, \quad 0 \le n_a \le 10 \tag{3.16}$$

The expected loss of losing power for a residential customer per day is \$100. Customers' loss L is calculated by the sum of customers' loss one each day during the entire recovery process. In the total T days of recovery, the total of customers' loss is given by Eq. (3.17):

$$L(n_a) = 100 \times \{3000T(n_a) - [3 \times 100 + n_a f_e(n_a)][1 + 2 + \cdots + (T(n_a) - 1)]\} \tag{3.17}$$

Adding customers' loss and the cost of crews and logistics yields the total cost of recovery C_t, which is given by Eq. (3.18):

$$C_t(n_a) = [f_c(n_a) + 5,000] T(n_a) + L(n_a) \tag{3.18}$$

The change of total cost relative to the change of the number of additional crews is depicted in Fig. 3.11. It can be observed that the optimal number of additional crews is 7 in this example.

After identifying the optimal number of additional crews, the error in estimating the optimal number of additional crew can be calculated as Eq. (3.19):

$$\varepsilon(\%) = \frac{n_e - n_{optm}}{n_{optm}} \times 100 \tag{3.19}$$

where n_e represents the estimated number of additional crews; n_{optm} represents the optimal number of additional crews. Then, the extra cost of recovery is calculated by the difference between the real cost and the minimum cost. Figure 3.12 displays the change in the extra cost of recovery relative to the change of the error in estimating the number of additional
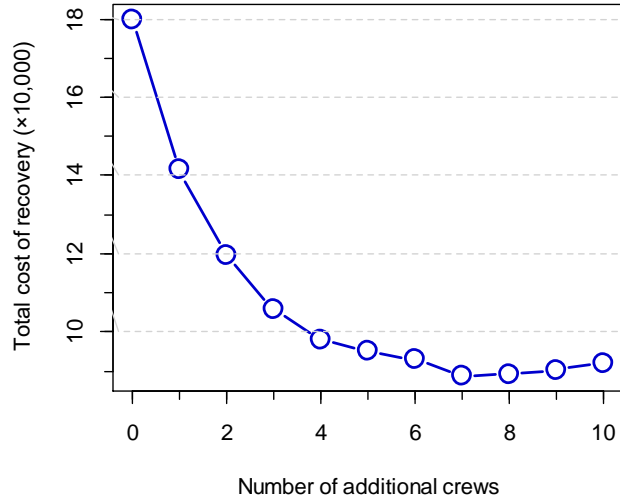
Figure 3.11: The relationship between the total cost of recovery and the number of additional crews

crews. Note that the negative error means underestimation while the positive error means overestimation. It can be observed from Fig. 3.12 that the accurate estimation of the number of additional crews has a great influence over the total cost of recovery: If the number is underestimated by around 40%, the extra cost will be approximately $10,000. Therefore, the ability to accurately predict the recovery rate can help minimize the overall cost of recovery by dynamically estimating the number of recovery crews needed as a function of time after a storm.
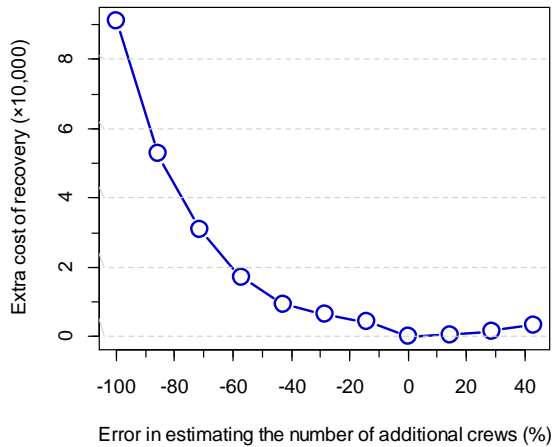
Figure 3.12: The relationship between the extra cost of recovery and the error in estimating the number of additional crews

The accurate estimation of recovery rate is also of great concern to decision-makers in government agencies because such disasters can cause tremendous economic loss and great inconvenience to residents and even heavy casualties in the affected areas. For example, Hurricane Ike caused a power outage that lasted from a few days to several weeks, affecting more than 2.8 million customers in the Greater Houston Area [119]. During Hurricane Irma, the dysfunction of the air conditioning system of a nursing home in Florida caused by the prolonged power failure led to the death of eight elderly [120].

3.5   Summary

Natural hazards can severely damage infrastructure systems and disrupt the activities of the customers they serve. In the management of infrastructure systems, the accurate estimation of the recovery rate is important to help responders and planners optimize the resource allocation during the recovery process and planning investments to improve the resilience against future disruptions. In this chapter, a new model, the hierarchical Bayesian kernel model, is proposed to predict the recovery rate of communities when data are lacking. This method addresses the problem of data scarcity by borrowing information from other similar subsets of the data and improves the estimation of parameters by assigning a distribution to the prior parameter that is updated using the entire dataset. The HBKM is compared to HBRM, the classical GLM, and the null model using recovery data on power outages from five of the largest storms that have hit Shelby County in Tennessee since 2007. HBKM yields the highest out-of-sample predictive accuracy on average. While HBRM achieves the highest frequency of yielding the highest log-likelihood and lowest RMSE, the proposed HBKM in this work requires less computation time while offering higher overall predictive accuracy as the error distribution of HBKM has a shorter tail compared to the error distribution of HBRM.

Chapter 4

Scalable Inference in Hierarchical Bayesian Model for Grouped Count Data

4.1   Introduction

Estimating the state of infrastructure systems, which is critical to infrastructure management in the aftermath of human-induced or natural hazards, often involves count data [**?**]. For example, power outage rate is commonly calculated with the number of customers who have regain power in a particular area after severe storms [65]. In gas or water distribution systems, the estimation of pipe repair rate requires the number of pipe breaks within a time period [121]. Hierarchical Bayesian Poisson regression models (HBPRMs) can be employed to analyze count data [122], especially when the dataset comes in groups. As a hierarchical Bayesian approach, HBPRMs offer the flexibility to capture the complex hierarchical structure of the count data, thereby improving the estimation accuracy of parameters for each group (e.g., areas, experiments, and events) [123, 124]. More specifically, this class of model can account for the individual- and group-level variations in estimating the parameters of interest and the uncertainty around the estimation of hyperparameters [125, 126, 75]. However, the flexibility of hierarchical models to capture the complex interactions in the data comes with a high computational cost because all the model parameters need to be estimated jointly [44, 127]. Furthermore, in big data settings, the dataset may be structured in many levels or groups [128], thus the hierarchical model may involve a large number of parameters to learn, which further increases the computational load. Consequently, HBPRMs demand more efficient inference algorithms.

By far the most popular algorithm for parameter estimation in hierarchical Bayesian (or in general Bayesian) model is Markov chain Monte Carlo (MCMC). MCMC algorithms obtain samples from a target distribution by constructing a Markov chain (irreducible and aperiodic) in the parameter space that has precisely the target distribution as its stationary dis-

tribution. This class of algorithms provide a powerful tool to obtain posterior samples when the exact full posterior distributions are only known up to a constant and direct sampling is not possible. However, one drawback of standard MCMC algorithms, such as Metropolis-Hastings algorithm (MH), is that they suffer from slow mixing [129, 130, 131], requiring a large number of Monte Carlo samples that grow with the dimension and complexity of the dataset [132]. Since practitioners often do not have researchers' luxury of running an MCMC chain for days to collect a large enough number of posterior samples [133], the application of standard MCMC algorithms to learn HBPRMs or other hierarchical Bayesian models for large datasets is significantly limited.

In remedy of the issue above, we propose an approximate Gibbs sampler to specifically infer HBPRMs in a computationally efficient way. The basic idea is to replace complex distributions that lead to a computational bottleneck with an analytical approximation based on conjugacy, such that sampling from the approximate distributions is significantly faster than from the original. Recent studies that use analytical approximation by exploiting conjugacy to accelerate MCMC-based inference in hierarchical Bayesian models include Refs. [127, 134]. Dutta et al. [127] use conjugacy to improve inference efficiency in large and complex hierarchical models. They show that the approximation by making use of conjugacy can be utilized even though the original hierarchical model is not fully conjugate. As an example in their study, the approximate full conditional distributions are derived when the likelihood function is a gamma distribution while the prior for the parameters are assumed to be multivariate normal and inverse Wishart distribution respectively. Berman [134] derives a Gaussian approximation to the conditional distribution of the normal random effects in hierarchical Bayesian binomial model (HBBM) using Taylor series expansion, such that Gibbs sampling can be applied to infer the HBBM more efficiently. Another study similar in spirit but not about hierarchical model is conducted by Chan and Vasconcelos [135] in which the data likelihood is approximated by Gaussian distribution to allow for faster inference of regression parameters in the Bayesian Poisson model for

crowd counting.

In this chapter, we focus specifically on accelerating the inference in HBPRMs for count data. The contributions of this chapter are twofold. First, we derive the closed-form approximation to the complex conditional posterior of the parameters and develop the approximate Gibbs sampler (AGS) algorithm to efficiently infer a general HBPRM using the approximate Markov chain while maintaining the inference accuracy. Therefore, the AGS enables one to use HBPRMs in large-scale applications. Second, we conduct multiple numerical experiments to demonstrate the performance of AGS against the start-of-the-art sampling algorithm.

4.2   Methodology

4.2.1   Hierarchical Bayesian Poisson Regression Model

Without loss of generosity, we consider a general HBPRM in which the coefficient for each covariate varies with the group (Eq. (4.1) to Eq. (4.5)). This model is applicable to count datasets in which the counts can be divided into multiple groups based on the covariates. Let $\mathscr{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$ be the dataset where $\boldsymbol{X} \in \mathbb{R}^{N \times K}$, $\boldsymbol{y} \in \mathbb{Z}_*^N$ where Z_* denotes the set of non-negative integers. An HBPRM assumes that each count, y_{ij}, follows a Poisson distribution. Then the log of each count is linearly linked to the covariates. In the hierarchical Bayesian paradigm, each of the parameters (regression coefficients) in the linear function follows a prior distribution with hyperparameter(s). A prior distribution is additionally specified on the hyperparameter(s) referred to as the hyperprior. Note that the hyperprior is shared among the parameters for the same covariate of all groups, thereby inducing dependency among the parameters that are dependent on the hyperprior [75]. When the variance of the hyperprior is reduced to zero, the hierarchical model is reduced to a non-hierarchical model. The mathematical formulation of the HBPRM is

$$y_{ij}|\lambda_{ij} \sim \text{Pois}(\lambda_{ij}), \text{ for } i = 1,...,n_j, \ j = 1,...,J, \tag{4.1}$$

66

$$\log \lambda_{ij} = \sum_{k=1}^{K} w_{jk} x_{ijk}, \text{ for } i = 1, ..., n_j, \ j = 1, ..., J, \ k = 1, ..., K, \tag{4.2}$$

$$w_{jk} | \mu_{w_k}, \sigma_{w_k}^2 \sim N(\mu_{w_k}, \sigma_{w_k}^2), \text{ for } j = 1, ..., J, \ k = 1, ..., K, \tag{4.3}$$

$$\mu_{w_k} | m, \tau^2 \sim N(m, \tau^2), \text{ for } k = 1, ..., K, \tag{4.4}$$

$$\sigma_{w_k}^2 | a, b \sim \text{IG}(a, b), \text{ for } k = 1, ..., K. \tag{4.5}$$

where y_{ij} is the i-th count within group j with an estimated mean of λ_{ij}. n_j is the number of data points in group j, w_{jk} is the regression coefficient of covariate k, x_{ijk} the i-th value in group j of covariate k. The prior for the coefficient of each covariate, μ_{w_k}, is assumed to be a Gaussian distribution (N) while the prior for the variance, $\sigma_{w_k}^2$, is assumed to be an inverse-gamma distribution (IG). The Gaussian and inverse-gamma distribution are specified as the prior for the mean and the variance respectively, such that we can exploit conditional conjugacy for analytical and computational convenience.

4.2.2   Inference

Give an HBPRM for analyzing an observed count dataset structured with multiple groups, $\mathcal{D}$, the task is to estimate the joint posterior density distribution of all the parameters, which is only known up to a constant. Writing

$$\Theta = \left\{ w_{11}, ..., w_{jk}, ..., w_{JK}; \mu_{w_1}, ..., \mu_{w_K}; \sigma_{w_1}^2, ..., \sigma_{w_K}^2 \right\},$$

the joint posterior factorizes as

$$p(\Theta | \boldsymbol{y}, \boldsymbol{X}) \propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} \text{Pois} \left(y_{ij} | \sum_{k=1}^{K} x_{ijk} w_{jk} \right) \prod_{k=1}^{K} N \left(w_{jk} | \mu_{w_k}, \sigma_{w_k}^2 \right) N \left(\mu_{w_k} | m, \tau^2 \right) \text{IG} \left(\sigma_{w_k}^2 | a, b \right). \tag{4.6}$$

However, sampling from the joint posterior is hardly an easy task because it does not admit a closed-form expression. Although the MH algorithm can be applied, the need to judiciously tune the step size for the desired acceptance rate often repel users from using this

algorithm [136, 137]. In comparison, Gibbs sampler is more efficient and does not require any tuning of the proposal distribution, therefore it has been used for Bayesian inference in a wide range of applications [138, 139]. Classical Gibbs sampling requires a conditional conjugate posterior distribution of each parameter (or block of parameters). The full conditional posteriors for implementing the Gibbs sampler are

$$p\left(w_{jk}|-\right) \propto \prod_{i=1}^{n_j} \text{Pois}\left(y_{ij}|\sum_{k=1}^{K} x_{ijk}w_{jk}\right) N\left(w_{jk}|\mu_{w_k}, \sigma_{w_k}^2\right), \tag{4.7}$$

$$p\left(\mu_{w_k}|-\right) \propto N\left(w_{1k},...,w_{Jk}|\mu_{w_k}, \sigma_{w_k}^2\right) N\left(\mu_{w_k}|m, \tau^2\right), \tag{4.8}$$

$$p\left(\sigma_{w_k}^2|-\right) \propto N\left(w_{1k},...,w_{Jk}|\mu_{w_k}, \sigma_{w_k}^2\right) \text{IG}\left(\sigma_{w_k}^2|\frac{a}{2}, \frac{b}{2}\right). \tag{4.9}$$

where $p\left(\cdot|-\right)$ represents the conditional posterior of a parameter of interest given the remaining parameters and the data. Due to the Gaussian-Gaussian and Gaussian-inverse-gamma conjugacy, Eq. (4.8) and Eq. (4.9) have an analytical expression [140]

$$p\left(\mu_{w_k}|-\right) \propto N\left(\mu_{w_k}\left|\frac{1}{\frac{1}{\tau^2} + \frac{J}{\sigma_{w_k}^2}}\left(\frac{m}{\tau^2} + \frac{\sum_{j=1}^{J} w_{jk}}{\sigma_{w_k}^2}\right), \frac{1}{\frac{1}{\tau^2} + \frac{J}{\sigma_{w_k}^2}}\right.\right) \tag{4.10}$$

$$p\left(\sigma_{w_k}^2|-\right) \propto \text{IG}\left(\sigma_{w_k}^2\left|\frac{a+J}{2}, \frac{b+\sum_{j=1}^{J}\left(w_{jk}-\mu_{wk}\right)^2}{2}\right.\right) \tag{4.11}$$

However, Eq. (4.7) does not admit an analytical solution because the Poisson likelihood is not conjugate to the Gaussian prior. One may resort to other algorithms, such as slice sampling [141], adaptive-rejection sampling [142], and Metropolis-within-Gibbs algorithm [143], to infer $p\left(w_{jk}|-\right)$, but those algorithms introduce an additional computational cost. Instead, we employ a Gaussian approximation to the Poisson likelihood given in Eq. (4.7) to obtain a closed-form solution to the conditional posterior of coefficients. With the closed-form solution, the complex inference of regression coefficients can be simplified to save computational resources. Reducing the computational cost of sampling from $p\left(w_{jk}|-\right)$

is critical for datasets with a large number of groups because the number of regression coefficients, $J \times K$, can be significantly larger than the number of prior parameters, $2K$.

4.2.3 Gaussian Approximation to Log-gamma Distribution

Consider a gamma random variable z with probability density function (pdf) given by

$$p(z|\alpha, \beta) = \frac{z^{\alpha-1} e^{-\frac{z}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}, \; \alpha > 0, \; \beta > 0, \tag{4.12}$$

where $\Gamma(\cdot)$ is the gamma function. α and β are the location parameter and the scale parameter respectively. Then the random variable, $\ln z \in R$, follows a log-gamma distribution. The mean and variance of log-gamma distribution are calculated by Eq. (4.13) and Eq. (4.14) respectively [144], which are given as

$$\mu_z = \psi_0(\alpha) + \ln\beta \tag{4.13a}$$

$$= -\gamma + \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+\alpha-1} \right) + \ln\beta \tag{4.13b}$$

$$= -\gamma + \sum_{n=1}^{\alpha-1} \frac{1}{n} + \ln\beta, \tag{4.13c}$$

$$\sigma_z^2 = \psi_1(\alpha) \tag{4.14a}$$

$$= \sum_{n=0}^{\infty} \frac{1}{(\alpha+n)^2} \tag{4.14b}$$

$$= \frac{\pi^2}{6} - \sum_{n=1}^{\alpha-1} \frac{1}{n^2}. \tag{4.14c}$$

In Eq. (4.13a) and Eq. (4.14a), $\psi_0(\cdot)$ and $\psi_1(\cdot)$ are the zeroth and the first order of polygamma functions respectively [145]. In Eq. (4.13b) and Eq. (4.13c), γ is the Euler-Mascheroni constant [146].

For large values of α, log-gamma distribution can be approximated by a Gaussian distribution [135, 147], i.e.

$$\ln z | \alpha, \beta \sim N\left(\psi_0\left(\alpha\right) + \ln\beta, \psi_1\left(\alpha\right)\right). \tag{4.15}$$

Setting $\alpha = y \in \mathbb{Z}_+$ and $\beta = 1$, Eq. (4.15) becomes

$$\ln z | y, 1 \sim N\left(\psi_0\left(y\right), \psi_1\left(y\right)\right). \tag{4.16}$$

Similarly, plugging in $\alpha = y$ and $\beta = 1$ and $\Gamma(n) = (n-1)!$ for $n \in \mathbb{Z}_+$, Eq. (4.12) becomes

$$p\left(z | y, 1\right) = \frac{z^{y-1}e^{-z}}{(y-1)!}. \tag{4.17}$$

Using the "change of variable" method and substituting $\ln z$ with v, we obtain the pdf of v [135]

$$p\left(v | y, 1\right) = p\left(z = e^v | y, 1\right) \frac{\partial e^v}{\partial v} \tag{4.18}$$

$$= \frac{1}{(y-1)!} e^{vy} e^{-e^v}. \tag{4.19}$$

Using Eq. (4.16) yields

$$\frac{1}{(y-1)!} e^{vy} e^{-e^v} \approx N\left(\psi_0\left(y\right), \psi_1\left(y\right)\right). \tag{4.20}$$

4.2.4 Closed-form Approximate Conditional Posterior Distribution

In the conditional posterior of coefficient w_{jk} given by Eq. (4.7), the likelihood function is

$$\prod_{i=1}^{n_j} \text{Pois}\left(y_{ij} | \sum_{k=1}^{K} w_{jk} x_{ijk}\right) = \prod_{i=1}^{n_j} \frac{1}{y_{ij}(y_{ij}-1)!} e^{v_{ij}y_{ij}} e^{-e^{v_{ij}}}. \tag{4.21}$$

Applying the approximation given by Eq. (4.20) yields

$$\prod_{i=1}^{n_j} \text{Pois}\left(y_{ij} | \sum_{k=1}^{K} w_{jk} x_{ijk}\right) \approx \prod_{i=1}^{n_j} \frac{1}{y_{ij}} N\left(\psi_0\left(y_{ij}\right), \psi_1\left(y_{ij}\right)\right). \tag{4.22}$$

Plugging in Eq. (4.22) into Eq. (4.7) we get

70

$$p\left(w_{jk}|-\right) \propto \exp\left[\frac{\left(w_{jk}-\mu_{w_k}\right)^2}{-2\sigma_{w_k}^2}\right]\prod_{i=1}^{n_j}\exp\left\{\frac{\left[\sum\limits_{k=1}^{K}x_{ijk}w_{jk}-\psi_0\left(y_{ij}\right)\right]^2}{-2\psi_1\left(y_{ij}\right)}\right\} \qquad (4.23)$$

$$= \exp\left\{\frac{\left(w_{jk}-\mu_{w_k}\right)^2}{-2\sigma_{w_k}^2}+\sum_{i=1}^{n_j}\frac{\left[\sum\limits_{k=1}^{K}x_{ijk}w_{jk}-\psi_0\left(y_{ij}\right)\right]^2}{-2\psi_1\left(y_{ij}\right)}\right\}. \qquad (4.24)$$

As the product of two Gaussians is still Gaussian, the posterior can also be written as

$$p\left(w_{jk}|-\right) \propto \exp\left[\frac{\left(w_{jk}-\widehat{\mu}_{w_{jk}}\right)^2}{\widehat{\sigma}_{w_{jk}}^2}\right], \qquad (4.25)$$

where $\widehat{\mu}_{w_{jk}}$ and $\widehat{\sigma}_{w_{jk}}^2$ are the mean and variance of the approximate Gaussian posterior.

Now we employ the "completing the squares" technique to find $\widehat{\mu}_{w_{jk}}$ and $\widehat{\sigma}_{w_{jk}}^2$. Note that terms that do not impact w_{jk} are regarded as a constant, i.e. C_i ($i = 1,...,4$) in the following equations. We start by writing Eq. (4.23) as

$$p\left(w_{jk}|-\right) \propto \exp\left\{\frac{\left(w_{jk}-\mu_{w_k}\right)^2}{-2\sigma_{w_k}^2}+\sum_{i=1}^{n_j}\frac{\left[x_{ijk}w_{jk}+\sum\limits_{h=1,h\neq k}^{K}x_{ijh}w_{jh}-\psi_0\left(y_{ij}\right)\right]^2}{-2\psi_1\left(y_{ij}\right)}\right\}. \qquad (4.26)$$

Let A be the exponent in Eq. (4.26) and we have

$$A = \frac{\left(w_{jk}-\mu_{w_k}\right)^2}{-2\sigma_{w_k}^2}+\sum_{i=1}^{n_j}\frac{\left[x_{ijk}w_{jk}+\sum\limits_{h=1,h\neq k}^{K}x_{ijh}w_{jh}-\psi_0\left(y_{ij}\right)\right]^2}{-2\psi_1\left(y_{ij}\right)} \qquad (4.27)$$

$$= \frac{\left(w_{jk}-\mu_{w_k}\right)^2\sum\limits_{i=1}^{n_j}\psi_1\left(y_{ij}\right)+\sigma_{w_k}^2\sum\limits_{i=1}^{n_j}\left[x_{ijk}w_{jk}+\sum\limits_{h=1,h\neq k}^{K}x_{ijh}w_{jh}-\psi_0\left(y_{ij}\right)\right]^2}{-2\sigma_{w_k}^2\sum\limits_{i=1}^{n_j}\psi_1\left(y_{ij}\right)}. \qquad (4.28)$$

Expanding the squares yields

$$A = \frac{\sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right) w_{jk}^2 - 2\mu_{w_k} \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right) w_{jk} + \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right) \mu_{w_k}^2}{-2\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)} +$$

$$\frac{\sigma_{w_j}^2 \sum\limits_{i=1}^{n_j} \left[x_{ijk}^2 w_{jk}^2 - 2x_{ijk} \psi_0\left(y_{ij}\right) w_{jk} + 2x_{ijk} \sum\limits_{h=1,h\neq k}^{K} x_{ijh} w_{jh} w_{jk} + C_1 \right]}{-2\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)}. \qquad (4.29)$$

Grouping the terms that impact w_{jk}^2 and w_{jk} respectively, we have

$$A = \frac{\left[\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} x_{ijk}^2 + \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right) \right] w_{jk}^2}{-2\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)} +$$

$$\frac{-2\left\{ \mu_{w_k} \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right) + \sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} x_{ijk} \left[\psi_0\left(y_{ij}\right) - \sum\limits_{h=1,h\neq k}^{K} x_{ijh} w_{jh} \right] \right\} w_{jk} + C_2}{-2\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)}. \qquad (4.30)$$

Dividing the numerator and denominator by the coefficient of the quadratic term, we get

$$A = \frac{w_{jk}^2 - 2\dfrac{\mu_{w_k} \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right) + \sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} x_{ijk} \left[\psi_0\left(y_{ij}\right) - \sum\limits_{h=1,h\neq k}^{K} x_{ijh} w_{jh} \right]}{\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} x_{ijk}^2 + \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)} w_{jk} + C_3}{-2\dfrac{\sigma_{w_j}^2 \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)}{\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} x_{ijk}^2 + \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)}} \qquad (4.31)$$

$$= \frac{\left\{ w_{jk} - \dfrac{\mu_{w_k} \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right) + \sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} x_{ijk} \left[\psi_0\left(y_{ij}\right) - \sum\limits_{h=1,h\neq k}^{K} x_{ijh} w_{jh} \right]}{\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} x_{ijk}^2 + \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)} \right\}^2}{-2\dfrac{\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)}{\sigma_{w_k}^2 \sum\limits_{i=1}^{n_j} x_{ijk}^2 + \sum\limits_{i=1}^{n_j} \psi_1\left(y_{ij}\right)}} + C_4. \qquad (4.32)$$

Using Eq. (4.32), Eq. (4.26) becomes

$$p\left(w_{jk}|-\right) \propto \exp\left(\frac{\left\{w_{jk} - \dfrac{\mu_{w_k}\sum_{i=1}^{n_j}\psi_1\left(y_{ij}\right)+\sigma_{w_k}^2\sum_{i=1}^{n_j}x_{ijk}\left[\psi_0\left(y_{ij}\right)-\sum_{h=1,h\neq k}^{K}x_{ijh}w_{jh}\right]}{\sigma_{w_k}^2\sum_{i=1}^{n_j}x_{ijk}^2+\sum_{i=1}^{n_j}\psi_1\left(y_{ij}\right)}\right\}^2}{-2\dfrac{\sigma_{w_k}^2\sum_{i=1}^{n_j}\psi_1\left(y_{ij}\right)}{\sigma_{w_k}^2\sum_{i=1}^{n_j}x_{ijk}^2+\sum_{i=1}^{n_j}\psi_1\left(y_{ij}\right)}}\right). \quad (4.33)$$

Completing the squares we get

$$\widehat{\mu}_{w_{jk}} = \frac{\mu_{w_k}+\sigma_{w_k}^2\sum_{i=1}^{n_j}\dfrac{x_{ijk}}{\psi_1\left(y_{ij}\right)}\left[\psi_0\left(y_{ij}\right)-\sum_{h=1,h\neq k}^{K}x_{ijh}w_{jh}\right]}{\sigma_{w_k}^2\sum_{i=1}^{n_j}\dfrac{x_{ijk}^2}{\psi_1\left(y_{ij}\right)}+1}, \quad (4.34)$$

$$\widehat{\sigma}_{w_{jk}}^2 = \frac{\sigma_{w_k}^2}{\sigma_{w_k}^2\sum_{i=1}^{n_j}\dfrac{x_{ijk}^2}{\psi_1\left(y_{ij}\right)}+1}. \quad (4.35)$$

Now that the full conditional posterior distributions can be expressed analytically, we can construct the approximate Gibbs sampler (Algorithm 2) to obtain posterior samples of the parameters in HBPRM efficiently.

Algorithm 2: Approximate Gibbs sampler

Input: X, y, # of samples as warmup N_0, # of desired samples N_1.
Output: Desired posterior samples, $\mu_k^{(i)}, \sigma_k^{2(i)}, w_{jk}^{(i)}$, for $i = 1,...,(N_0+N_1)$.

 1: Define initial sample, $\mu_k^{(0)}, \sigma_k^{2(0)}, w_{jk}^{(0)}$.
 2: **for** $i = 2$ to $N_0 + N_1$ **do**
 3:     **for** $k = 1$ to K **do**
 4:         Sample hyperparameter $\mu_k^{(i)}$ according to Eq. (4.10).
 5:         Sample hyperparameter $\sigma_k^{2(i)}$ according to Eq. (4.11).
 6:         **for** $j = 1$ to J **do**
 7:             Sample each parameter $w_{jk}^{(i)}$ according to Eq. (4.25).
 8:         **end for**
 9:     **end for**
10: **end for**

4.3 Experiments

To illustrate the proposed AGS, we compare the performance of AGS and NUTS under different real count datasets. Details about each of these datasets are presented below.

4.3.1 Data Description

Synthetic data: The synthetic datasets are generated according to the following model:

$$x_1 \sim \text{Unif}(0.1, 2), \tag{4.36a}$$

$$x_2 \sim \text{Unif}(0.1, 1), \tag{4.36b}$$

$$x_3 \sim \text{Unif}(0.1, 0.5), \tag{4.36c}$$

$$x_4 \sim \text{Unif}(1, 10), \tag{4.36d}$$

$$x_5 \sim \text{Unif}(0.5, 5), \tag{4.36e}$$

$$x_6 \sim \text{Unif}(10, 100), \tag{4.36f}$$

$$n_j \sim \text{Unif}\left(10^4, 10^6\right), \ j = 1, ..., J \tag{4.36g}$$

$$w_{jk} \sim \mathcal{N}(0.001, 0.001), \ j = 1, 2, ..., J, k = 1, ..., K \tag{4.36h}$$

$$y_{ij} = \left(e^{w^T X}\right)_{\text{norm}} n_j, \ j = 1, ..., J. \tag{4.36i}$$

where $(\cdot)_{\text{norm}}$ represents min-max normalizing function. We generated 15 synthetic datasets (S1, ..., S15) with varying number of data points N_d, number of covariates N_c, and number of groups N_g (see Table 4.1).

Power outage data: The power outage data includes 11 datasets (denoted by P1,..., P11) that were recorded after 11 disruptive events. Covariates in each dataset include precipitable water vapor (kg·m^{-2}), 10-meter northward wind speed (m/s), 10-meter eastward wind speed (m/s), surface pressure (Pa), time of measurement (h). The response variable is the outage count at each time of measurement following each of the disruptive events.

Covid-19 test data: The Covid-19 test dataset is obtained from Ref. [148]. This dataset

includes the days, t, after exposure to Covid-19 in each study. The response variable is the number of patients who tested positive. The total number of test cases is 380 while the number of studies after removing the zero count is 298. The test cases are grouped by studies and the total number of studies is 11. Following Ref. [148], the exposure is assumed to have occurred five days before symptom onset and $\log(t)$, $\log(t)^2$, $\log(t)^3$ are used as the covariates.

Bike sharing data: The bike sharing data are available from UC Irvine Machine Learning Repository [149, 150]. The bike sharing data include daily bike rental counts for 729 days, normalized temperature, normalized humidity, and casual bike rentals. The daily dataset is grouped by whether or not the bike rental occurs in a working day.

4.3.2 Experimental Setup

In the HBPRM for the count datasets listed above, we employ $N(0,1)$ [100] and $\mathrm{IG}(1,1)$ as a weakly-informative prior for μ_w and σ_w^2 respectively. The competing algorithm considered is No-U-Turn sampler (NUTS), an extension to Hamiltonian Monte Carlo (HMC) algorithm [129] that frees users from tuning the proposals [151]. In the numerical experiments, NUTS is implemented with Stan [100, 152]. Numbers are averaged over 4 runs of 10000 iterations for each algorithm, discarding the first 5000 samples as warm-up (burn-in). We compare AGS with NUTS in terms of average sampling time per 1000 iterations (T_s), sampling efficiency (E_s), R^2, and Root Mean Square Error (RMSE). Sampling efficiency is quantified as the mean effective sampler size (ESS) over the average sampling time per 1000 iterations. ESS is calculated according to Ref. [75, Chapter 11]. All the experiments are implemented with R (version 3.6.1) [153] on a Windows 10 desktop computer with a 3.40 GHz Intel Core i7-6700 CPU and 16.0 GB RAM.

Table 4.1: Performance of NUTS and AGS under different simulated datasets

Dataset	Characteristics			T_s (s)		E_s		R^2		RMSE	
	N_d	N_c	N_g	NUTS	AGS	NUTS	AGS	NUTS	AGS	NUTS	AGS
S1	200	2	10	1.51	**0.96**	**649.28**	33.46	0.9390	**0.9450**	6500	**6191**
S2	400	2	10	2.66	**0.98**	**373.46**	34.99	0.9430	**0.9500**	5823	**5455**
S3	800	2	20	5.05	**1.63**	**195.21**	18.63	0.9576	**0.9626**	3526	**3310**
S4	200	3	10	5.76	**1.25**	**170.39**	5.28	0.9614	**0.9660**	4235	**3976**
S5	400	3	10	16.63	**1.29**	**59.78**	2.03	0.9683	**0.9710**	3450	**3305**
S6	800	3	20	40.53	**2.44**	**24.30**	1.80	0.9677	**0.9719**	3993	**3728**
S7	200	4	10	7.64	**1.63**	**129.25**	1.59	0.9716	**0.9760**	2955	**2718**
S8	400	4	10	23.72	**1.94**	**41.69**	1.50	0.9639	**0.9701**	4406	**4012**
S9	800	4	20	45.00	**3.31**	**21.99**	0.49	0.9740	**0.9770**	2958	**2790**
S10	200	5	10	12.12	**2.06**	**81.26**	0.70	0.9574	**0.9631**	4941	**4593**
S11	400	5	10	24.70	**2.41**	**39.68**	0.49	0.9782	**0.9826**	2514	**2245**
S12	800	5	20	64.00	**4.12**	**15.44**	0.27	0.9767	**0.9804**	3446	**3157**
S13	200	6	10	14.66	**2.44**	**67.49**	0.23	0.9817	**0.9876**	2720	**2721**
S14	400	6	10	42.01	**2.63**	**20.17**	0.24	0.9922	**0.9948**	1339	**1128**
S15	800	6	20	93.20	**4.97**	**8.44**	0.17	0.9910	**0.9940**	1994	**1629**

4.3.3 Results

The performance of NUTS and AGS under different datasets are summarized in Table 4.1 (simulated datasets) and Table 4.2 (real datasets). It can be observed that under both the simulated and real datasets AGS outperforms NUTS in the average sampling time, especially when the size of the datasets is large. Depending on the dataset, AGS can offer one to two orders of magnitude speedup of inference efficiency against NUTS while achieving competitive inference accuracy measured by R^2 and RMSE. However, AGS suffers from low ESS, therefore the sampling efficiency of AGS is significantly lower than that of NUTS in almost all the datasets, even though the sampling speed of AGS is much faster than NUTS.

As the scalability of algorithms is crucial for large-scale hierarchical data, we show the average sampling time of the two algorithms for different sized simulated and real datasets in Fig. 4.1. Empirically, the sampling time of both samplers grows with the size of dataset, but when compared to NUTS, the increase in the sampling time of AGS is significantly lower, suggesting great scalability.

Table 4.2: Performance of NUTS and AGS under different real datasets

Dataset	Characteristics			T_s (s)		E_s		R^2		RMSE	
	N_d	N_c	N_g	NUTS	AGS	NUTS	AGS	NUTS	AGS	NUTS	AGS
P1	3817	5	56	885.65	**15.76**	1.12	**1.25**	0.9730	**0.9801**	13558	**11621**
P2	2467	5	50	652.87	**13.71**	**1.50**	0.62	0.9873	**0.9884**	1446	**1384**
P3	1548	5	35	387.47	**9.41**	**2.54**	0.92	0.9850	**0.9870**	1974	**1833**
P4	632	5	26	165.73	**6.67**	**5.94**	0.46	**0.9923**	0.9918	**1327**	1373
P5	520	5	16	135.16	**4.31**	**7.27**	1.85	0.9934	**0.9940**	3473	**3312**
P6	421	5	17	118.73	**4.54**	**8.25**	2.68	0.9908	**0.9918**	2526	**2387**
P7	375	5	23	39.86	**5.76**	**24.75**	2.41	**0.9459**	0.9355	**3574**	3903
P8	247	5	10	8.75	**2.78**	**111.91**	2.38	**0.9744**	0.9729	**795**	818
P9	157	5	8	5.48	**2.24**	**179.87**	4.07	0.9964	**0.9967**	803	**766**
P10	115	5	6	4.49	**1.72**	**218.17**	6.75	0.9869	**0.9915**	7715	**6222**
P11	63	5	4	5.49	**1.25**	**177.38**	0.92	**0.9356**	0.9027	**251**	308
Bike share	729	3	2	9.09	**0.98**	**109.54**	24.75	**0.6743**	0.6292	**1101**	1175
Covid test	298	3	11	34.60	**2.59**	**28.57**	18.54	0.8517	**0.8582**	2.53	**2.47**
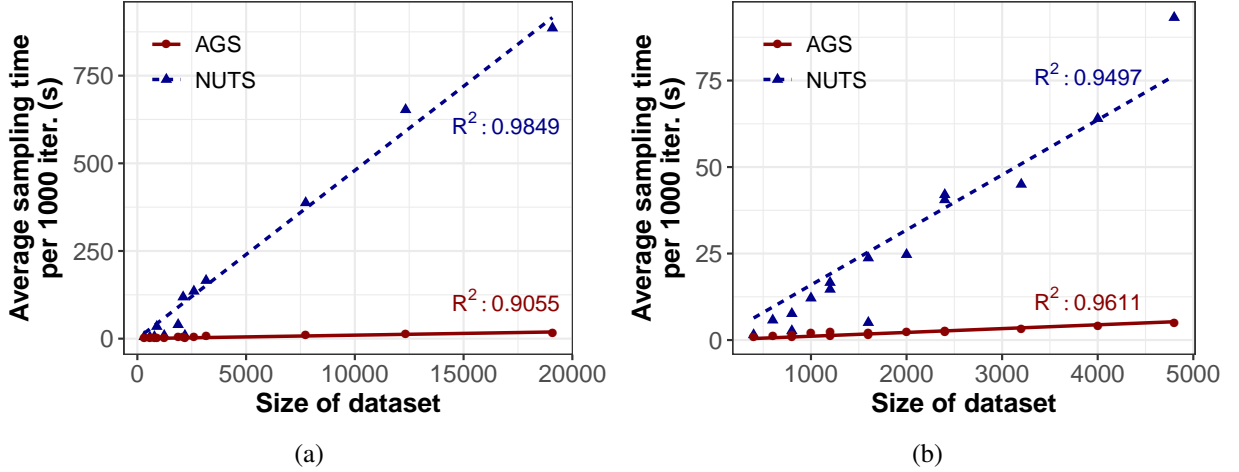
Figure 4.1: Scalability of NUTS and AGS on (a) real datasets and (b) simulated datasets. The size of a dataset is calculated by $N_d \times N_c$.

4.4   Summary

in this chapter, we propose a scalable approximate Gibbs sampling algorithm for the HBPRM for grouped count data, which can be applied to estimate the system recovery state. Our algorithm builds on the approximation of data-likelihood with Gaussian distribution such that the conditional posterior for coefficients have a close-form solution. Empirical evaluations show that the proposed algorithm outperforms the state-of-the-art sam-

pling algorithm, NUTS, in inference efficiency. The improvement in efficiency is greater for larger datasets, suggesting good scalability. Due in part to the Gibbs updates, the AGS trades off greater accuracy for slower mixing Markov chains, leading to a much lower effective sample size and therefore lower sampling efficiency [154]. However, when sampling time is of great concern to model users, AGS might be the only feasible option. As the approximation quality improves with larger counts, our algorithm works better for count datasets in which the counts are large, such as power recovery datasets.

Chapter 5

Modeling Interdependencies of Infrastructure Systems Using Stochastic Block Models

5.1   Introduction

Modern society depends on critical infrastructure for security, safety, health, and over-all well-being. Recent changes in social and economic development as well as advances in technology have led to increasing interdependent connections between critical infrastructure systems [155]. On one hand, these interdependencies can enhance the overall efficiency and robustness of these systems under normal operations by providing a platform for information, services, and product sharing. On the other hand, highly connected critical infrastructure can be more vulnerable to hazards whereby a disruptive event in one system can result in cascading failures across other connected infrastructure systems [11, 12, 13]. As such, interdependencies among infrastructures must be taken into account to understand the operational characteristics of infrastructure systems [156] and to inform operations and design of robust and resilient systems. One of the major challenges in evaluating the impact of disruptive events on IISs lies in understanding the influence of multiple interdependencies on the systemic performance after a disruption.

Rinaldi et al. [157] define interdependency as "the bidirectional relationship between two infrastructures through which the state of each infrastructure influences or is correlated to the state of the other." After formalizing the concept of interdependency, Rinaldi et al. [157] propose four principal classes of interdependencies: 1) physical interdependency, when energy or materials flow between two systems; 2) cyber interdependency, when information is transmitted between systems; 3) geographic interdependency, when the state of one system can be altered due to spatial proximity of another system; 4) logical interdependency, when two systems influence each other via a mechanism that is not based on physical, cyber, or geographic connection. Although other classifications have also been

proposed [158, 159, 160, 161], the study in this chapter employs the classification proposed by [157] as the proposed approach can be directly extended to other types of interdependencies [21] which can be found in Ref. [8]. Modeling the resilience or recovery of IISs has been extensively studied [162, 160, 163, 164, 165, 166, 167, 168]. In most studies, IISs are often modeled in a deterministic way by establishing interdependency links first and then evaluating the systemic performance. These links are established between randomly generated nodes [169], or based on either degree centrality [165] or spatial proximity (minimum Euclidean distance) [170, 164, 171, 21, 30]. One approach to assessing resilience under uncertainty is the use of Bayesian networks (BNs), such as Refs. [172, 173, 174]. The majority of studies that use BNs examine a single network instead of multiple interdependent networks. Furthermore, BNs are an acyclic graphical model and thus have great difficulty in handling the bi-directional relationship in interdependent networks.

The majority of existing studies do not consider the uncertainty associated with the interdependency between the networks and few studies can be found in the literature that address dynamic and uncertain interdependencies. Full knowledge of interdependencies across infrastructure networks, despite its key importance, is often not available due to the lack of data [175]. Therefore the topology of IISs is subject to uncertainty. Neglect or improper characterization of this uncertainty can lead to underestimation or overestimation of system performance as the metrics cannot be assessed within an acceptable level of fidelity [176, 17].

The objective of this chapter is to model the uncertainty and dynamics of interdependencies after the disruption in order to guide restoration strategies. The specific contributions of this chapter are listed below.

1. The application of statistical network models to evaluate uncertain network interdependencies. Specifically, the use of stochastic block models provides a probabilistic characterization of interdependency links between infrastructure networks.

2. A formulation for estimating the likelihood that an interdependency link exists using

multiple predictors to represent major factors influencing the presence of interdependency links. To avoid zero values in normalizing the variables that will be used in the denominator, a variant of min-max normalization called *truncated min-max normalization* is developed.

3. A network recovery strategy based on the dynamic ranking of component importance. Network restoration based on dynamic component ranking results in a faster recovery of IISs when compared to static ranking.

The remainder of this chapter is outlined as follows: In Section **??**, we introduce the mathematical representation of stochastic block model (SBM). Section 5.2 outlines the approach to assessing the resilience of IISs along with two restoration strategies based on static and dynamic component importance ranking. In Section 5.3, the proposed method is illustrated using interdependent water and power networks. Finally, Section 5.4 provides the conclusion along with a discussion for future work.

5.2 Methodology

5.2.1 Estimation of Interdependency Between Networks

The study in this chapter assumes that the interdependency links connect the demand node of one network to a supply node in another network. This assumption is realistic for many real-world IISs. For example, in interdependent water and power network, the interdependency links are established as follows:

1. The end-user node (demand node) in the water network and the power station (supply node) in the power network that requires water for generating steam.

2. The end-user node (demand node) in the power network and the pumping station (supply node) in the water network that requires electric power [21].

In order to model the uncertainty of interdependencies across IISs, interdependency links are estimated probabilistically. The model shown in Eq. (5.1) is proposed to evaluate the probability an interdependency link exists based on a set of nodal attributes. These attributes can represent the physical, economic, and social characteristics of the networks. The attributes considered in this chapter account for the distance between the networks, the number of customers served by the nodes, and the vulnerability of the customers represented with the Social Vulnerability Index (SoVI). Distance is used as a predictor as distance-based features are found to be significant in estimating the missing link [177]. The model includes the number of customers (modeled using the population) since disrupted components serving a large number of customers might be given higher priority during the restoration. The third predictor, SoVI, was originally developed to identify the characteristics of the population that render social communities vulnerable to external disturbances [178]. It is calculated based on a large number of factors, including socioeconomic status, age, house type, education level, race, among others. SoVI values range from 0 to 1 with higher values indicating a higher level of vulnerability. Social vulnerability can inform underlying and intangible interdependency links that would guide restoration activities to achieve community resilience by prioritizing vulnerable customers who might disproportionately suffer more damage from disruptions [179, 180].

$$P(A_{ij} = 1) = \beta_0 + \beta_1 \times d_{v_i v_j}^{-1} + \beta_2 \times p_{v_{nd}} + \beta_3 \times s_{v_{nd}} \tag{5.1}$$

In Eq. (5.1), $A_{ij} = 1$ indicates the presence of an interdependency link between v_i and v_j (v_i and v_j belong to different individual networks). β_i ($i = 1, 2, 3$) represents the regression coefficient, $p_{v_{nd}}$ represents the population served by the parent node of the interdependency link on which the dependent (child) node relies, and $s_{v_{nd}}$ represents the SoVI of the census tract where the parent node is located. Note that both $d_{v_i v_j}$ and $p_{v_{nd}}$ should be normalized before they are used in the model and the obtained probabilities should also be normalized such that they sum to one. The model can be conveniently refined when data about real-

world interdependency links and nodal attributes are available.

Since the reciprocal of the distance between nodes is used in Eq. (5.1), zero values must be avoided in the normalized distance. To this end, the min-max normalization is modified (Eq. (5.2)) to scale the distance data to the range $\left[\frac{\alpha}{1+\alpha}, 1\right]$. As $\frac{\alpha}{1+\alpha}$ should approach zero, a small value of α is preferred, i.e. $\alpha \ll 1$. Therefore, $\frac{\alpha}{1+\alpha} \approx \alpha$, indicating that the normalized data have an approximate lower bound equal to α.

$$x' = \frac{x - x_{\min} + \alpha \left(x_{\max} - x_{\min} \right)}{\left(1 + \alpha \right) \left(x_{\max} - x_{\min} \right)} \tag{5.2}$$

In Eq. (5.2), x represents the data to be normalized and x' represents data after normalization. This variant of min-max normalization can be referred to as *truncated min-max normalization*.

5.2.2 Component Importance Ranking and Restoration

In the study of this chapter, the sequence of infrastructure network restoration is determined according to the ranking of components. Resilience-based component importance ranking can help inform resource allocation and prioritization of repair activities when multiple components are damaged [69]. The importance is quantified by the relative resilience improvement of the interdependent networks after each component is restored individually. Two restoration sequences are proposed, the first one is based on *static ranking* and the second one is based on *dynamic ranking*. In the case of static ranking, the damaged components to be repaired are ranked only once before the start of recovery, therefore the benchmark for resilience improvement is the resilience of the IISs at the disrupted state $R(t_d)$. The importance of a component can be calculated using Eq. (5.3).

$$I^{c_i} = \frac{R^{c_i}(t_d) - R(t_d)}{R(t_d)} \tag{5.3}$$

In Eq. (5.3), I^{c_i} represents the importance of component i, $R^{c_i}(t_d)$ represents the resilience of the IISs after component i is restored at time t_d before initiating recovery activi-

ties. The damaged components are then restored sequentially according to the ranking.

In the case of dynamic ranking, the damaged components are ranked at every time step until all the components are restored. At the time t, the benchmark for resilience improvement is $R(t)$. Accordingly, the dynamic component importance can be computed using Eq. (5.4).

$$I^{c_i}(t) = \frac{R^{c_i}(t) - R(t)}{R(t)}, \, t \in (t_d, t_f) \tag{5.4}$$

In Eq. (5.4), $I^{c_i}(t)$ is the importance of component i at time t and $R^{c_i}(t)$ is the resilience of the IISs after component i is restored at time t. When $t = t_d$, Eq. (5.4) becomes Eq. (5.3), meaning that static importance ranking is simply the initial dynamic importance ranking of all the damaged components. The steps for dynamic component importance ranking, coupled with component restoration, are summarized in Algorithm 3.

Algorithm 3: Dynamic component importance ranking for resilience assessment

---
Input: Adjacency matrix, node type, node coordinates, component failure probability, etc.
Output: The resilience over the restoration process $R(t)$, $t \in (t_d, t_f)$.
 1: Compute the initial resilience $R(t_d)$.
 2: **for** $t = 1$ to T **do** $\triangleright$ T: time needed to restore all the damaged components
 3:     **for** $i = 1$ to N_t^r **do** $\triangleright$ N_t^r: # of remaining damaged components at time t
 4:         Rank the remaining components based on the importance calculated by Eq. (5.4).
 5:         Return the resilience after restoring the component with the highest importance, c_t.
 6:     **end for**
 7:     Remove the component c_t from the list of remaining components.
 8:     Calculate the resilience according to Eq. (5.5) and record it as $R(t)$.
 9: **end for**

5.2.3   Resilience Assessment of Interdependent Networks

Network performance is determined by the ratio of the number of functional demand nodes to the total number of demand nodes of each network. The resilience of the inter-dependent networks is calculated as the weighted average of the resilience of individual

networks, Eq. (5.5).

$$R(t) = \sum_{k=1}^{K} w_k \frac{n_t^k - n_d^k}{n_0^k - n_d^k}, \, t \in (t_d, t_f) \tag{5.5}$$

In Eq. (5.5), K is the number of infrastructure networks. w_k is the weight for individual network k with $\sum_{k=1}^{K} = 1$. n_t^k, n_d^k, and n_0^k represent the number of functional demand nodes of network k at time t, t_d, and t_0, respectively.

As a network component in real-world IISs can lose functionality due to (i) direct physical damage caused by the disruptive event or (ii) loss of necessary supply from the components of other networks, the following definitions are proposed to differentiate between the two cases.

Definition 5.2.1. *A network component is operable if the physical entity represented by the component is not damaged by the disruptive event.*

Definition 5.2.2. *A network component is functional if the component is operable and can receive supply from other components to maintain functionality.*

In a single network, we assume that an operable demand node is functional as long as it is connected to a functional supply node according to the two definitions, so the functionality of a demand node can be evaluated using shortest path algorithms, such as Dijkstra's algorithm [181], and Floyd-Warshall algorithm [95]. The Floyd-Warshall algorithm is used because it can find the shortest path between all pairs of nodes simultaneously. However, in interdependent networks, the functional supply node in an individual network, for example, G_1, must also be connected to a demand node in the other network G_2, as shown in Fig. 1.3. The following theorem shows the conditions that must be satisfied for a node to be functional in interdependent networks.

Proposition 5.2.1. *In undirected interdependent networks represented by the underlying graph G, a demand node v is functional if v is contained in an operable cycle that consists of at least one interdependency link or if v is connected to such a cycle.*

85

Proof of Proposition 5.2.1. Define that a path or cycle is operable if all the nodes and edges they contain are operable. For the sake of contradiction, assume that a demand node v_x of G_1 can be functional even if (a) no operable cycles that contains interdependency links exist in G, or (b) the demand node is neither a part of nor connected to such a cycle. In case (a), since v_x is functional, there must exist an operable path from v_x to a supply node in G_1. Let P be the longest functional path that starts from v_x. Since v_x is functional, the supply node closest to v_x in G_1 must be adjacent to a functional demand node in G_2. Following this logic, P must be made up of D-T-S paths that connect a demand node, transmission node, and supply node by sequence, such as $P_1, P_2, \ldots, P_n$ in Fig. 5.1. Note that by definition, repeated nodes are not allowed in P otherwise a cycle will be complete. Let v_y be the endpoint of the last D-T-S path in P. Because v_y is the endpoint, it does not receive supply from the demand node of another network, so v_y is not functional, contradicting that P is functional. In case (b), suppose a functional demand v_x is connected to some supply node v_z in G_1 and v_z is not part of any cycle that consists of an interdependency link. Using the same logic of case (a), v_z will not be functional, leading to a contradiction. □
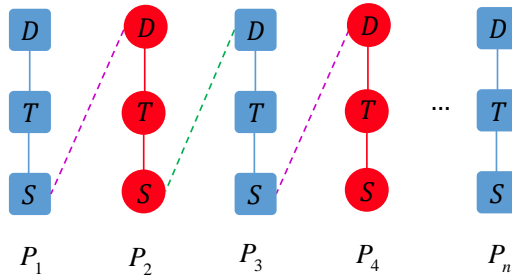
Figure 5.1: Assumed functional longest path in interdependent networks without cycles

Based on this *sufficient* condition, a key step to determine the functionality of a node is to detect functional cycles in the interdependent networks. Building on the *Depth First Search* (DFS) [182] algorithm for detecting cycles in graphs and the Floyd-Warshall algorithm for checking connectivity, the steps for detecting the functional nodes in interdependent networks are provided in Algorithm 4.

Algorithm 4: Identification of functional nodes

Input: Adjacency matrix A of the interdependent networks
Output: An array that contains the ID of functional nodes
1: Apply DFS to detect cycles.
2: Select cycles that contain at least one interdependency link (The start node and end node belong to different individual networks).
3: Apply the Floyd-Warshall algorithm to find the distance from all nodes to the nodes contained in the selected cycles.
4: If the distance value is greater than zero, then the corresponding node is functional; otherwise, the node is not functional.

Once the functional nodes are detected, the resilience of the interdependent networks at time t is obtained using Eq. (5.5). Considering the uncertainty of interdependency links and the failure of network components, the process for assessing the resilience of interdependent networks is proposed as follows:

1. Import data on the interdependent networks, including the adjacency matrix, distance between nodes (normalized), population of the census tract where each node is located (normalized), and SoVI, among others.

2. Estimate interdependency links based on Eq. (5.1).

3. Draw samples of interdependency links according to their respective probabilities.

4. Define disruption scenario and calculate the disruption intensity at the site of each network component.

5. Calculate failure probability of each network component (including the estimated interdependency links) using empirical equations from HAZUS, a standardized methodology provided by the Federal Emergency Management Agency (FEMA) to estimate potential losses from multiple types of hazards [4]. For simplicity, components are assumed to be inoperable once they are damaged, i.e., partial functionality is not considered.

6. Generate a sufficient number of network configurations, i.e., the possible network topology after the damage of a subset of components by the disruptive event e ac-

cording to different probabilities. The randomness of the network configurations can be modeled by first comparing a random vector $\boldsymbol{u}$ ($\boldsymbol{u} \sim U(0,1)$) to the vector containing the failure probability of each component, and then, if the failure probability of a certain component is greater than the random number drawn from $U(0,1)$, the component is assumed to be inoperable.

7. Generate new interdependency links if the links or nodes on the interdependency links are damaged. New interdependency links are considered to account for the interdependency that emerges in the aftermath of the disruption.

8. Rank components based on the static ranking or dynamic ranking (Algorithm 3).

9. Restore one component and record resilience at each time step t.

5.3 Case Study

5.3.1 Data Description

The system of two interdependent power and water networks shown in Fig. 5.2 is used to illustrate the proposed model in this chapter. The water distribution network includes six elevated storage tanks, nine pumping stations, thirty-four intermediate delivery nodes, and seventy-one water pipes while the power network (modified from Ref. [183]) consists of fourteen gate stations, twenty-three 23-kV substations and twenty-two 12-kV substations, respectively. Gate stations and pumping stations are considered to be the supply facilities, 23-kv substations and storage tanks as transmission facilities, and 12-kv substations and intermediate delivery nodes as demand facilities. It should be noted that intermediate delivery nodes are the intersection points of water pipes and they are assumed to be undamaged after earthquakes since there do not exist large-scale facilities at the site of these nodes. In this case study, physical interdependency is considered since pumping stations rely on the 12-kv substations for power supply [184] while the gate stations require clean water to generate high-pressure steam to drive the turbines. Note that the power stations

may rely on the river nearby instead of the water distribution network for water to be used for cooling purposes [171].

Data on the population and SoVI at the census tract level where each node of the interdependent networks is located are publicly available through the Census Bureau [185] and Centers for Disease Control and Prevention [186], respectively.
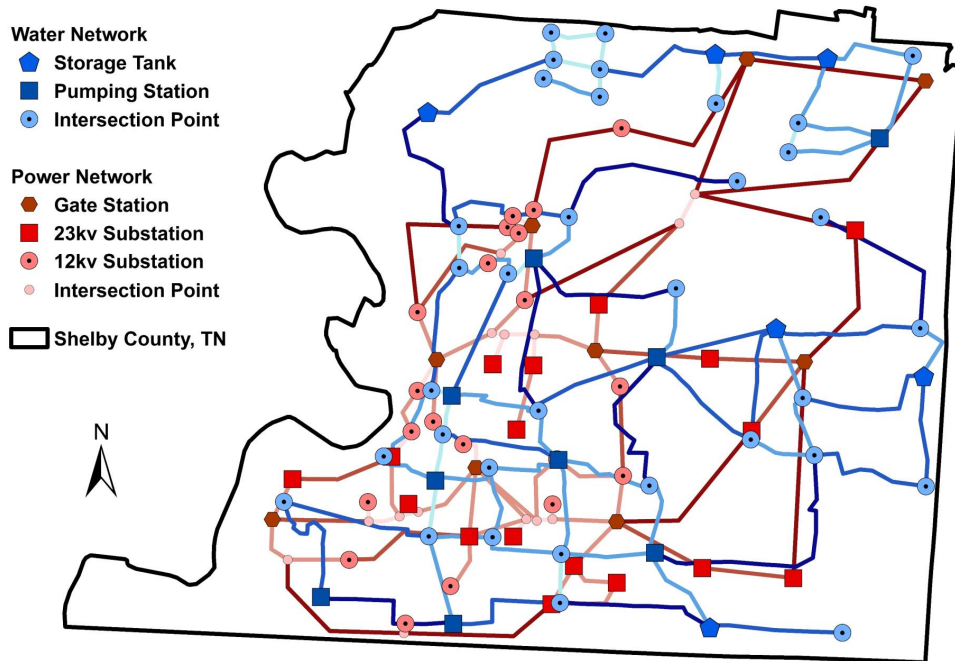
Figure 5.2: Interdependent water and power networks of Shelby County, TN

5.3.2   Results

5.3.2.1   Disruption Intensity and Component Fragility

The disruptive event considered in this case study is a hypothetical earthquake. An earthquake centered at $(35.3 \, N, \, 90.3 \, W)$ (the maximum probable earthquake [98]) is used to calculate the seismic intensity at the site of each component. In calculating the failure probability of each component given the earthquake scenario, the fragility curve of power and water network facilities under earthquakes is adapted from HAZUS. In HAZUS, five

damage states are defined: none (ds_1), minor (ds_2), moderate (ds_3), extensive (ds_4), and complete (ds_5). Each damage state corresponds to one component fragility curve. This case study adopts the damage state ds_5. More details on the determination of seismic intensities at the site of network components and the failure probability of components can be found in the Earthquake Model Technical Manual [187]. Components of the power network are assumed to respond in a similar way as water network to seismic events. This assumption can be relaxed when data about the fragility of power network components become available.

5.3.2.2   Interdependency Links

The estimated parameters of the probabilistic model for interdependency links from Eq. (5.1) are $\beta_0 = 0$, $\beta_1 = 0.5$, $\beta_2 = \beta_3 = 0.25$, $\alpha = 0.01$. For one particular node in each network, the probabilities of interdependency links between this node and components from the other network are shown in Fig. 5.3 and Fig. 5.4. Specifically, the likelihood of an interdependency link is described by the probability using the estimation model from Eq. (5.1) under two scenarios. The first scenario only considers the physical interdependence by including the distance as the sole predictor, while the second scenario incorporates social aspects such as population and SoVI besides the distance. It is noted that the interdependency links with the highest probability of occurrence are different under the two scenarios, especially in Fig. 5.4 where the two rankings of power nodes are entirely different. In addition, the distribution for these probabilities of all possible links becomes a bit flatter when social aspects are considered. The reason is that the ranking of nodes based on the population or SoVI is strikingly different from that based on the distance. The change in the probability of interdependency links after including the social attributes indicates that the social aspects do not differentiate between possible interdependency links like the geographical distance does. Collecting additional information allows for a more comprehensive model with additional predictors to identify the contributing factors to the

existence of an interdependency link.
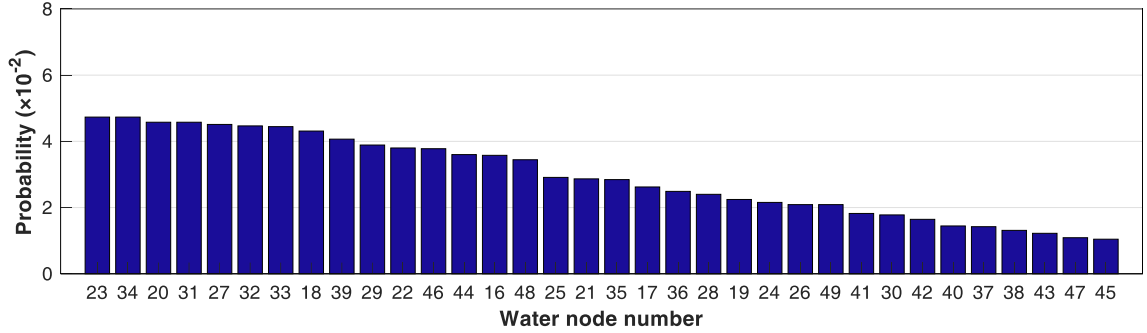
(a)
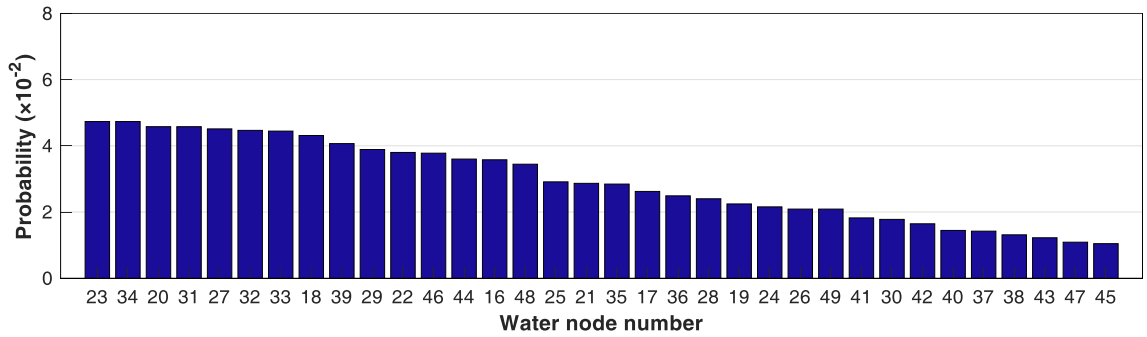
(b)

Figure 5.3: Probability of interdependency links between water nodes and gate stations. (a) With population and SoVI. (b) Without population and SoVI

5.3.2.3 Resilience Assessment

Without loss of generality, this chapter assumes that the water network and power network weigh equally in calculating the resilience of the interdependent networks, thus $w^k = \frac{1}{2}$ with $K = 2$. Due to the probabilistic failure of network components, each of the possible network structures generated by each estimation of interdependency links can have a myriad of new network structures at time t_d after the disruption. To characterize the randomness in the network structure at time t_d, 100 simulation runs are used to obtain the mean value and the range of resilience for each potential scenario of interdependency links between the networks. During the recovery process, it is assumed that one component can
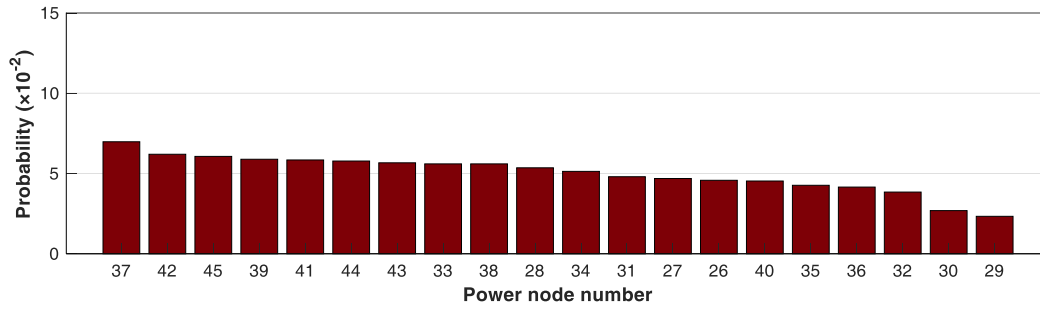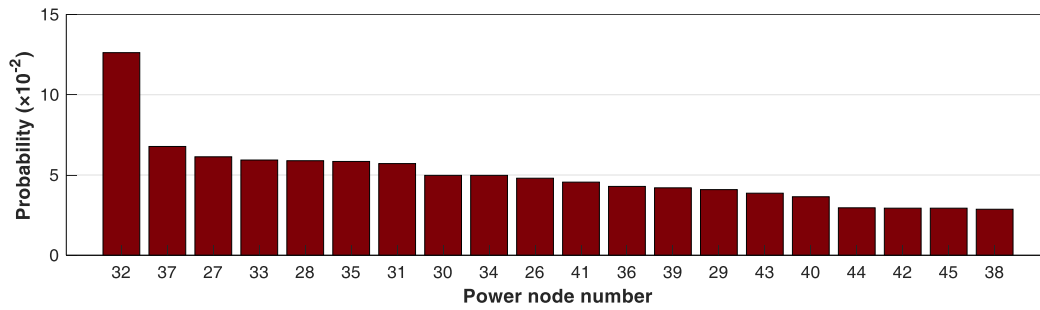
Figure 5.4: Probability of interdependency links between power nodes and pumping stations. (a) With population and SoVI. (b) Without population and SoVI

be restored at each time step, which can be modified to represent other possible restoration strategies. The resilience curves from the disrupted stage to the new stable stage under different seismic intensities (Fig. 5.5) show the response of different network structures to the same disruptive event.

Throughout the recovery process, the mean value, lower bound, and upper bound of resilience based on dynamic ranking is greater than or equal to those based on static ranking. Dynamic ranking yields a more rapid recovery process and improved resilience in the early stage of restoration. Further, the overall range under the static ranking approach is much larger than the dynamic ranking approach. This outcome suggests that accounting for the dynamic nature of interdependencies by updating the ranking of components at each time step given the new configuration of the network is critical to improving the resilience of these systems.
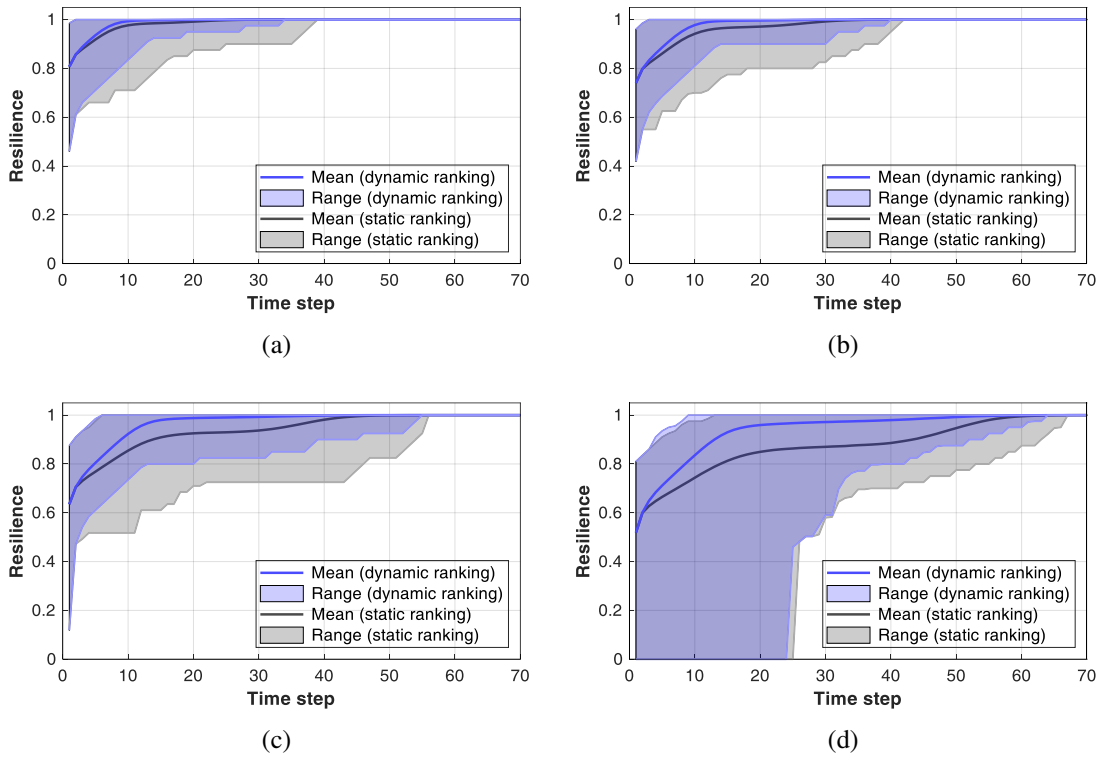
Figure 5.5: Resilience curves with lower/upper bounds under different seismic intensities. (a) Magnitude=5.0. (b) Magnitude=6.0. (c) Magnitude=7.0. (d) Magnitude=8.0.

5.4   Summary

The interdependency links across infrastructure networks are uncertain and can change over time in response to disruption, restoration activities, and reconfiguration of infrastructure networks. As such, modeling the resilience of IISs should account for the inherent uncertainty and dynamic behavior of interdependency links after a disruption. The outcome of this work is a data-driven stochastic method that estimates the likelihood that an interdependency link exists based on the behavior of the networks after a disruption. The calculation of the probability of an interdependency link incorporates multiple predictors describing geographic proximity, physical connection, and social impact. It is therefore possible to predict how interdependency links emerge and disappear during and after a disruption. Given this additional information on infrastructure network behavior, restoration

strategies can be adapted in real time following a dynamic ranking process for the most important components. In addition, this approach provides the ability to identify factors that impact the existence and strength of an interdependency link. This work expands on existing studies of IISs where only geographic and physical interdependence are considered [170, 171, 30]. For example, the incorporation of SoVI in the estimation of interdependency links and consequent dynamic ranking of components significantly improves the resilience of IISs by decreasing the restoration time after a disruption. The restoration activities in this case are informed by the updated prioritization of components based on the needs of the communities measured here by SoVI.

This chapter presents a new approach founded in Stochastic Block Models to capture the uncertainty associated with interdependency links of IISs. The approach is flexible that it can be generalized to other types of IISs and the probability of interdependency links can incorporate additional and different factors and variables. For example, in order to capture other sources of uncertainty such as the success of the restoration or secondary failures, the probability of an interdependency link or the reconfiguration of the networks can be dependent on stochastic variables describing these phenomena. The results from the case study on the interdependent water and power networks of Shelby County, Tennessee show that estimation of the presence of interdependency links based on the distance between network components can be significantly different from the estimation based on distance, population, and SoVI. This work demonstrates the importance of evaluating uncertain interdependency links by comparing two restoration strategies, one based on dynamic ranking given updates on the dynamic behavior of interdependencies and the other is static. The recovery based on dynamic component importance ranking results in faster restoration and improved resilience.

Chapter 6

Two-stage Stochastic Programming Approach for Integrated Protection and Restoration of
Interdependent Infrastructure Networks

6.1   Introduction

Extreme events and disasters pose a tremendous risk on the stability, security, and prosperity of a nation's economy, particularly because of their severe adverse impact on critical infrastructures (CIs). Recent enormous property, life, and economic loss incurred by hurricanes, such as Hurricane Irma and Hurricane Harvey in 2017, are among the prime examples. Due to the vulnerability of CIs against extreme events, there is increasing emphasis on enhancing the safety and resilience (recoverability) of CIs to hedge risks against the largely uncertain disruptions [188, 189, 190, 191].

The restoration of infrastructure networks in disruptive situations can be improved through network hardening or reconfiguration before the occurrence of extreme events. Large-scale network reconfiguration is usually not an option due to the prohibitive cost and excessively long time required [192]. Further, designing an efficient network hardening strategy can be a challenge as well because the perfect information about extreme events is almost impossible to come by since those extreme events are random by nature and empirical data about those events are often lacking. Thus, in order to improve the restoration under a limited budget, it is critical to integrate the protection and restoration of IISs.

Integrating protection and restoration decisions for IISs has drawn considerable attention from scholars and has been demonstrated to enhance the resilience of infrastructure networks [189, 19, 20, 21, 22].In particular, Nurre et al. [19] consider the integrated network design and restoration scheduling problem following an extreme event. The decision variable including components to install into a network and scheduling the installation to maximize the cumulative weighted network flow over the time horizon of planning.

In a companion study, Cavdaroglu et al. [20] extend the model to optimize the restoration planning and scheduling decisions of IISs with the objective of minimizing the total cost incurred over the horizon. Considering multiple types of interdependencies, González et al. [21] introduce the interdependent network design problem for optimal reconstruction strategy. However, the scheduling of repair crews and the uncertainty around the disruption are not taken into account. Considering the balance between the protection and restoration before and after a disruption, Ghorbani-Renani et al. [22] propose a tri-level protection-interdiction-restoration model to enhance the resilience of multiple interdependent networks for the worst-case scenario of disruption.

With the exception of recent work by Fang and Zio [189], few recent studies have considered the uncertainty around disruptions on the IISs. Fang and Zio [189] propose a two-stage adaptive two-stage robust optimization model for resilience enhancement considering the most-likely worse-case scenario of network failure. In their study, the first stage problem incorporates protection decision variables while the second stage considers decision variables concerning component restoration state, supply, and network flow. However, the scheduling of repair crews is not incorporated and a very limited number of scenarios are considered.

In this chapter, a stochastic programming approach is proposed to optimize the restoration of a coupled power-gas network. Specifically, a two-stage, mixed-integer stochastic program is developed with the objective of minimizing the total cost of protection before the disruption and post-disruption restoration cost. The stochastic model can characterize the uncertainty around the impact of disruptive events from possible multiple hazard types. The first stage decisions seek to provide the optimal protection strategy of the coupled networks while the second stage decisions are made to enhance the recoverability of the couple networks. In the proposed program, interdependencies are modeled through linking variables and logical constraints. We consider interdependent power and gas networks (IPGNs) in formulating the model because IPGNs have been increasingly important in the

U.S. during recent years due to the robust supply of low-priced natural gas, increased efficiency, high operational flexibility and low carbon footprint [193, 194, 195]. In IPGNs, compressors in the natural gas network may depend on the power grid for electricity while the power grid requires natural gas to fuel the gas-fired power generators [196].

The stochastic programming approach has been widely applied to planning and restoration management of IISs under uncertainty. Typically, the sample average approximation method is applied to obtain an approximate solution [197] with a reasonable level of accuracy. However, as the number of scenarios increases, the problem can be computationally challenging to solve because the number of second-stage decision variables can be very large even for interdependent networks of modest size. The solution methods used in the aforementioned studies, such as the tailored covering decomposition in Ref. [22] and the adapted nested cutting plane decomposition in Ref. [189], may not be applicable to networks of large scale since those methods have only been tested on cases with limited worse-case scenarios in the respective studies. In this chapter, we integrate meta-heuristics and iterative regularization techniques [198] to solve the problem under a significantly large number of scenarios.

The contributions of this section are threefold. First, we formulate a two-stage stochastic mixed-integer program to identify the optimal decisions on the protection and restoration of IISs under uncertainty around the disruptive events. Unlike Ref. [189], we also integrate decisions on the scheduling of repair crews in the second stage. Therefore, our model can be employed to assist managers in utility companies in designing the optimal schedule for repair crews. Note that although we illustrate the proposed methodology on interdependent power and gas networks, this framework can be applied to some other types of IISs, such as interdependent power-water networks. Second, we propose solution methods based on meta-heuristics and iterative regularization that can provide high quality solution to the two-stage mixed-integer program within a reasonable amount of time. Third, the computational results presented in the computational experiments demonstrate the value of the

solution obtained by the stochastic program.

6.2 Integrated Protection and Restoration of Interdependent Infrastructure Systems

This section presents a two-stage mixed-integer stochastic program for the protection and restoration of IISs under uncertain disruptions. We refer to the associated problem as the *integrated protection and restoration of interdependent infrastructure networks* (IPRIN) problem.

6.2.1 Problem description

IISs are model as a set of networks (e.g., Fig. 1.3) consisting of nodes (vertices) and links (directed edges). The movement of resources is modeled as network flow along the links. Formally, IISs $\mathcal{M}$ consist of a set of $|\mathcal{M}|$ different networks. Each network $m \in \mathcal{M}$ is represented by a directed graph $\mathcal{G} = (\mathcal{N}^m, \mathcal{A}^m)$, where $\mathcal{N}^m$ and $\mathcal{A}^m$ denote the node set and arc set of network m, respectively. The nodes can be categorized into supply nodes $\mathcal{N}_s$, transmission nodes $\mathcal{N}_t$, and demand nodes $\mathcal{N}_d$. The arc set $\mathcal{A}^{m \to l}$ represents network l's dependence on network m. Each component (node or link) is associated with a capacity, protection cost, restoration cost, and restoration time. Since the components in IISs are subject to damage from disruptive events, utility companies that run the IISs need to fortify the components before the disruptive event (first stage) and restore the damaged components after the disruption (second stage) in order to minimize the total cost incurred during the restoration. Specifically, first stage decisions (a.k.a there-and-now decisions) involves choosing the components to be hardened under the budget limit B, which can be modeled with a binary variable x. For example, x_i^m indicates that node $i \in \mathcal{N}^m$ is chosen to be hardened. The realization of the disruption is modeled with a binary variable β indicating whether or not the respective component is damaged by the disruption. After the disruption, restoration of the damaged IISs is initiated to repair the damaged components so that the service can be restored as soon as possible. In this stage, the decisions are to

98

jointly plan the supply at each supply node q, the flow on each arc (f for intra-links or h for interdependent links), and the schedule of repair crews (v). The finite planning horizon is divided into discrete time periods, given by the set $\mathcal{T}$. However, due to the randomness of disruptions, decision-makers do not have information about the disruption scenario. As such, the second stage decisions are not made until the disruption occurs and the second stage decisions are coupled with the first-stage decisions. Typically, the randomness about the disruption can be modeled with a large number of scenarios generated using Monte Carlo simulation of disruption. Note that when the first-stage decision variables, referred to as complicating variables, are fixed, the problem naturally decomposes into scenarios [43]. Now the problem is to minimize the expected total cost over all possible scenarios. This problem can be formulated as a two-stage mixed-integer stochastic program with complete mixed-integer recourse because the second-stage problem is feasible for all possible first-stage decisions [199, 200].

6.2.2   Problem Assumptions

The underlying assumptions in the model for the IPRIN problem are listed below.

1. Network components are damaged at the end of the first stage and restoration is initiated immediately at the beginning of the second stage.

2. Hardened components will be protected from the damage.

3. A repair crew can only work on a single disrupted component at a time. The repair crew has to work continuously on one component until it is fully repaired.

4. Damaged components will not be functional until they are fully repaired.

5. The direction of power flow can be reversed during the restoration.

6. Electric power generators can also receive supply from other power generators.

6.2.3 Notations

The following notations are used to facilitate the introduction of the proposed two-stage stochastic program.

Indices and index sets

$\mathcal{M}$	the set of networks.
$\mathcal{T}$	the set of time periods, $\mathcal{T} = \{1, \ldots, T\}$.
$\mathcal{N}^m$	the set of nodes in network m.
$\mathcal{A}^m$	the set of arcs in network m, $k = (i, j) \in \mathcal{A}, i, j \in \mathcal{N}$.
$\mathcal{A}^{m \to l}$	the set of dependency arcs, $m, l \in \mathcal{M}$.
Ω	the set of scenarios.

Parameters

u_k^m	arc capacity of arc $k \in \mathcal{A}^m$.
u_k^{ml}	arc capacity of arc $k \in \mathcal{A}^{m \to l}$.
d_{it}^m	demand of node $i \in \mathcal{N}^m$ at time $t \in \mathcal{T}$
$\bar{q}_i^m$	supply capacity of node $i \in \mathcal{N}^m$.
$\tau_i^{r,m}$	required recovery time of node $i \in \mathcal{N}^m$.
$\tau_k^{r,m}$	required recovery time of arc $k \in \mathcal{A}^m$.
$c_i^{r,m}$	cost per time period to repair node $i \in \mathcal{N}^m$.
$c_k^{r,m}$	cost per time period to repair arc $k \in \mathcal{A}^m$.
$c_{it}^{s,m}$	cost per unit for unmatched demand at node $i \in \mathcal{N}^m$ at time $t \in \mathcal{T}$.
$c_k^{f,m}$	cost per unit for flow on arc $k \in \mathcal{A}^m$.
$c_i^{h,m}$	hardening cost for node $i \in \mathcal{N}^m$.
$c_k^{h,m}$	hardening cost for arc $k \in \mathcal{A}^m$.
B	total budget allocated for hardening.
R^m	the number of repair crews.

p^{ω} probability of scenario $\omega \in \Omega$ occurring.

$\beta_i^{m\omega}$ $\begin{cases} 1 & \text{if node } i \in \mathcal{N}^m \text{ fails under scenario } \omega \in \Omega, \\ \\ 0 & \text{otherwise.} \end{cases}$.

$\beta_k^{m\omega}$ $\begin{cases} 1 & \text{if arc } k \in \mathcal{A}^m \text{ fails under scenario } \omega \in \Omega, \\ \\ 0 & \text{otherwise.} \end{cases}$.

α^{ml} flow conversion rate from network m to l.

Decision variables

x_i^m indicator whether node $i \in \mathcal{N}^m$ is hardened.

x_k^m indicator whether arc $k \in \mathcal{A}^m$ is hardened.

$y_{it}^{m\omega}$ indicator whether a repair crew start repairing at node $i \in \mathcal{N}^m$ at time $t \in \mathcal{T}$.

               under scenario $\omega \in \Omega$.

$y_{kt}^{m\omega}$ indicator whether a repair crew starts repairing at arc $k \in \mathcal{A}^m$ at time $t \in \mathcal{T}$

               under scenario $\omega \in \Omega$.

$z_{it}^{m\omega}$ indicator whether node $i \in \mathcal{N}^m$ functions at time $t \in \mathcal{T} \cup \{0\}$ under scenario $\omega \in \Omega$.

$z_{kt}^{m\omega}$ indicator whether arc $k \in \mathcal{A}^m$ functions at time $t \in \mathcal{T} \cup \{0\}$ under scenario $\omega \in \Omega$.

$v_{it}^{m\omega}$ indicator whether a repair crew is working at node $i \in \mathcal{N}^m$ at time $t \in \mathcal{T}$

               under scenario $\omega \in \Omega$.

$v_{kt}^{m\omega}$ indicator whether a repair crew is working at arc $k \in \mathcal{A}^m$ at time $t \in \mathcal{T}$

               under scenario $\omega \in \Omega$.

$w_{it}^{m\omega}$ indicator whether node $i \in \mathcal{N}^m$ is repaired by time $t \in \mathcal{T}$ under scenario $\omega \in \Omega$.

$w_{kt}^{m\omega}$ indicator whether arc $k \in \mathcal{A}^m$ is repaired by time $t \in \mathcal{T}$ under scenario $\omega \in \Omega$.

$f_{kt}^{m\omega}$ flow on arc $k \in \mathcal{A}^m$ at time $t \in \mathcal{T}$ under scenario $\omega \in \Omega$.

$h_{kt}^{m \to l, \omega}$ flow on arc $k \in \mathcal{A}^{m \to l}$ at time $t \in \mathcal{T}$ under scenario $\omega \in \Omega$.

$q_{it}^{m\omega}$ supply at node $i \in \mathcal{N}^m$ at time $t \in \mathcal{T}$ under scenario $\omega \in \Omega$.

$s_{it}^{m\omega}$ unmet demand at node $i \in \mathcal{N}^m$ at time $t \in \mathcal{T}$ under scenario $\omega \in \Omega$.

6.2.4 Mathematical Model

The extensive formulation (deterministic equivalent) of the IPRIN based on the two-stage stochastic programming approach is presented as follows:

$$\min \quad \sum_{\omega \in \Omega} p^{\omega} \left(\sum_{m \in \mathcal{M}} \left\{ \sum_{t \in \mathcal{T}} \left[\sum_{i \in \mathcal{N}_d^m} c_{it}^{s,m} s_{it}^{m\omega} + \sum_{i \in \mathcal{N}^m} c_i^{r,m} v_{it}^{m\omega} + \sum_{k \in \mathcal{A}^m} \left(c_k^{f,m} f_{kt}^{m\omega} + c_k^{r,m} v_{kt}^{m\omega} \right) \right] + \right.\right.$$

$$\left.\left. \left[\sum_{i \in \mathcal{N}^m} c_i^{r,m} \tau_i^{r,m} (1 - z_{iT}^{m\omega}) + \sum_{k \in \mathcal{A}^m} c_k^{r,m} \tau_k^{r,m} (1 - z_{kT}^{m\omega}) \right] \right\} \right) \tag{6.1a}$$

$$\text{s.t.} \quad \sum_{m \in \mathcal{M}} \left(\sum_{i \in \mathcal{N}^m} c_i^{h,m} x_i^m + \sum_{k \in \mathcal{A}^m} c_k^{h,m} x_k^m \right) \le B \tag{6.1b}$$

$$\sum_{k=(i,j) \in \mathcal{A}^m} f_{kt}^{m\omega} - \sum_{k=(j,i) \in \mathcal{A}^m} f_{kt}^{m\omega} + \sum_{l \in \mathcal{M}} \sum_{k=(i,j) \in \mathcal{A}^{m \to l}} h_{kt}^{m \to l, \omega} = q_{it}^m + s_{it}^{m\omega} - d_i^m$$

$$\forall m \in \mathcal{M}, i \in \mathcal{N}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1c}$$

$$f_{kt}^{m\omega} \le u_k^m z_{kt}^{m\omega} \quad \forall m \in \mathcal{M}, k \in \mathcal{A}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1d}$$

$$f_{kt}^{m\omega} \le u_k^m z_{it}^{m\omega} \quad \forall m \in \mathcal{M}, k \in \mathcal{A}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1e}$$

$$h_{kt}^{m \to l, \omega} \le u_k^{m \to l} z_{it}^{m\omega} \quad \forall m, l \in \mathcal{M}, k = (i,j) \in \mathcal{A}^{m \to l}, t \in \mathcal{T}, \omega \in \Omega \tag{6.1f}$$

$$q_{it}^{m\omega} = \sum_{l \in \mathcal{M}} \sum_{k=(j,i) \in \mathcal{A}^{l \to m}} \alpha_{ml} h_{kt}^{l \to m, \omega}$$

$$\forall m \in \mathcal{M}, i \in \mathcal{N}^m \cap \{i' \mid \exists l \in \mathcal{M}, (j,i') \in \mathcal{A}^{l \to m}\}, t \in \mathcal{T}, \omega \in \Omega \tag{6.1g}$$

$$v_{it}^{m\omega} = \sum_{t'=t-\tau_i^{r,m}+1}^{t} y_{it'}^{m\omega} \quad \forall m \in \mathcal{M}, i \in \mathcal{N}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1h}$$

$$v_{kt}^{m\omega} = \sum_{t'=t-\tau_k^{r,m}+1}^{t} y_{kt'}^{m\omega} \quad \forall m \in \mathcal{M}, k \in \mathcal{A}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1i}$$

$$\sum_{i \in \mathcal{N}_m} v_{it}^{m\omega} + \sum_{k \in \mathcal{A}_m} v_{kt}^{m\omega} \le R^m \quad \forall m \in \mathcal{M}, t \in \mathcal{T}, \omega \in \Omega \tag{6.1j}$$

$$w_{it}^{m\omega} = \sum_{t'=1}^{t-\tau_i^{r,m}} y_{it'}^{m\omega} \quad \forall m \in \mathcal{M}, i \in \mathcal{N}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1k}$$

$$w_{kt}^{m\omega} = \sum_{t'=1}^{t-\tau_k^{r,m}} y_{kt'}^{m\omega} \quad \forall m \in \mathcal{M}, k \in \mathcal{A}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1l}$$

$$z_{it}^{m\omega} \leq z_{it-1}^{m\omega} + w_{it}^{m\omega} \quad \forall m \in \mathcal{M}, i \in \mathcal{N}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1m}$$

$$z_{kt}^{m\omega} \leq z_{kt-1}^{m\omega} + w_{kt}^{m\omega} \quad \forall m \in \mathcal{M}, i \in \mathcal{A}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1n}$$

$$z_{i0}^{m\omega} \leq (1 - \beta_i^{m\omega}) + x_i^m \quad \forall m \in \mathcal{M}, i \in \mathcal{N}^m, \omega \in \Omega \tag{6.1o}$$

$$z_{k0}^{m\omega} \leq (1 - \beta_k^{m\omega}) + x_k^m \quad \forall m \in \mathcal{M}, k \in \mathcal{A}^m, \omega \in \Omega \tag{6.1p}$$

$$z_{it}^{m\omega} \geq z_{it-1}^{m\omega} \quad \forall m \in \mathcal{M}, i \in \mathcal{N}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1q}$$

$$z_{kt}^{m\omega} \geq z_{kt-1}^{m\omega} \quad \forall m \in \mathcal{M}, i \in \mathcal{A}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1r}$$

$$0 \leq s_{it}^{m\omega} \leq d_{it}^m \quad \forall m \in \mathcal{M}, i \in \mathcal{N}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1s}$$

$$0 \leq q_{it}^{m\omega} \leq \bar{q}_i^m \quad \forall m \in \mathcal{M}, i \in \mathcal{N}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1t}$$

$$f_{kt}^{m\omega} \geq 0 \quad \forall m \in \mathcal{M}, k \in \mathcal{A}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1u}$$

$$h_{kt}^{m \rightarrow l, \omega} \geq 0 \quad \forall m \in \mathcal{M}, k \in \mathcal{A}^{m \rightarrow l}, t \in \mathcal{T}, \omega \in \Omega \tag{6.1v}$$

$$x_i^m \in \{0, 1\} \quad \forall m \in \mathcal{M}, i \in \mathcal{N}^m \tag{6.1w}$$

$$x_k^m \in \{0, 1\} \quad \forall m \in \mathcal{M}, k \in \mathcal{A}^m \tag{6.1x}$$

$$y_{it}^{m\omega}, z_{it}^{m\omega}, w_{it}^{m\omega}, v_{it}^{m\omega} \in \{0, 1\} \quad \forall m \in \mathcal{M}, i \in \mathcal{N}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1y}$$

$$y_{kt}^{m\omega}, z_{kt}^{m\omega}, w_{kt}^{m\omega}, v_{kt}^{m\omega} \in \{0, 1\} \quad \forall m \in \mathcal{M}, k \in \mathcal{A}^m, t \in \mathcal{T}, \omega \in \Omega \tag{6.1z}$$

In model (6.1), we minimize the expected total cost, including flow costs, repairing costs, slack penalty costs, and penalty on components being unrepaired at the end of the time horizon over all disruption scenarios. Constraint (6.1b) limits the first-stage cost within the budget. We set the budget to be smaller than the amount required to harden all components, in order to prevent trivial solutions. Constraint (6.1c) imposes the flow conservation for nodes. Constraints (6.1d) and (6.1e) relate the network flow decision variables with the functionality status of arcs. Constraint (6.1f) mandates that node j in a network l, which depends on node i in network m, will not be functional until node i is restored. Constraint (6.1g) couples different networks by flow conversion and ensures that the converted flow from network l to a supply node in network m is equal to the supply

at the node. The term $\mathcal{N}^m \cap \{i' \mid \exists l \in \mathcal{M}, (j, i') \in \mathcal{A}^{l \to m}\}$ represents the set of nodes in network m that depend on the supply from network l. Constraints (6.1h) and (6.1i) enforce that a component will not be operational until it is restored by the crew. Constraint (6.1j) caps the number of repair crews. Constraints (6.1k) and (6.1l) ensure that a component will be restored by a time period after a repair crew has worked on it for the required number of time periods. Constraints (6.1m) to (6.1n) ensures that a component will not be functional until it is repaired. Constraints (6.1o) and (6.1p) state the effect of hardening decisions and the disruption to the initial functionality status of components. That is, hardened components will remain functional after the disruption and components that are not hardened will not be functional if they are damaged by the disruptive event. Constraints (6.1q) to (6.1z) state the nature of the decision variables. The number of decision variables in model (6.1) is

$$\left\{ 2 \left(|\mathcal{N}| + |\mathcal{A}| \right) + \left(6|\mathcal{N}| + 2|\mathcal{A}| + |\mathcal{A}^{m \to l}| \right) |\mathcal{T}| \right\} |\mathcal{M}| |\Omega|$$

while the number of constraints is

$$1 + (|\mathcal{N}| + |\mathcal{A}|) |\mathcal{M}| (1 + |\Omega|) + \left\{ 1 + 2|\mathcal{A}^{m \to l}| + |\mathcal{N}_d^{l \to m}| + 11 \left(|\mathcal{N}| + |\mathcal{A}| \right) \right\} |\mathcal{M}| |\mathcal{T}| |\Omega|$$

where $\mathcal{N}_d^{l \to m}$ represents the term $\mathcal{N}^m \cap \{i' \mid \exists l \in \mathcal{M}, (j, i') \in \mathcal{A}^{l \to m}\}$.

6.3   Solution Technique

The proposed two-stage stochastic program can be computationally challenging to solve because the number of second-stage decision variables can be very large even for interdependent networks of modest size. In our experience, off-the-shelve solvers (e.g., Gurobi) can typically solve instances with less than 100 scenarios (limited by RAM size of a computer), but a practical instance can have thousands of scenarios. Therefore, we develop a novel approach based on the *iterative regularization* technique to obtain a near-optimal solution with a reasonable amount of time.

Algorithm 5: Random walk iterative regularization

Input: m, ρ, maximum number of iterations κ_{max}, and $|\Omega|$ failure scenarios.

Output: $\hat{g}$, the LUB on the optimal objective value of model (6.1) under $|\Omega|$ scenarios.

  1: Set $\hat{g}$ to a significantly large value.   ▷ Initialize $\hat{g}$

  2: Solve model (6.1) under m randomly chosen scenarios to obtain the current candidate solution $\boldsymbol{X}^*$ and let $\hat{\boldsymbol{X}} \leftarrow \boldsymbol{X}^*$.   ▷ Initialize $\hat{\boldsymbol{X}}$

  3: **while** $\kappa \le \kappa_{max}$ **do**

  4:       Solve model (6.2) under m randomly chosen scenarios to obtain $\boldsymbol{X}^*$ and let $\hat{\boldsymbol{X}} \leftarrow \boldsymbol{X}^*$.

  5:       Solve model (6.1) under $|\Omega|$ scenarios with $\boldsymbol{X}$ fixed to $\hat{\boldsymbol{X}}$ to obtain the optimal

  6:       objective value g^*.

  7:       **if** $g^* < \hat{g}$ **then**

  8:          $\hat{g} \leftarrow g^*$ ▷ Update the best-so-far objective value

  9:       **end if**

10:       $\kappa \leftarrow \kappa + 1$

11: **end while**

6.3.1 Iterative Regularization

Because a direct solution to the approximate problem in model (6.1) can still be challenging as the number of scenarios can be explosively large. We solve multiple small-sized problems to obtain plausible optimal solutions to the first-stage decisions $\hat{x}$, then solve the original problem under a large number of scenarios with the first stage decisions fixed to the most plausible optimal solution obtained, which gives an upper bound of the original model under a large number of scenarios. The lowest optimal objective value obtained by conducting this solution technique multiple times is the *least upper bound* (LUB). To exploit the solutions obtained from each smaller-size problem, the regularization technique is employed to impose a penalty on deviation from the previous candidate optimal solution to the first-stage problem [198]. The regularized model is given as

$$\min \quad (6.1\text{a}) + \rho \sum_{m \in \mathcal{M}} \left(\sum_{i \in \mathcal{N}^m} |x_i^m - \hat{x}_i^m| + \sum_{k \in \mathcal{A}^m} |x_k^m - \hat{x}_k^m| \right) \tag{6.2a}$$

$$\text{s.t.} \quad (6.1\text{b}) \text{ to } (6.1\text{z}) \tag{6.2b}$$

In model (6.2), $\rho > 0$ represents the regularization strength. A larger value of ρ indicates heavier regularization, leading to a smaller deviation of the new candidate solution from the current solution. Note that the constraints in model (6.2) are the same as model (6.1). The iterative regularization algorithm is given by Algorithm 5. To simplify the notation, we use $\hat{X}$ to represents the incumbent solution.

Algorithm 6: SA iterative regularization

Input: m, ρ, maximum number of iterations κ_{max}, $|\Omega|$ failure scenarios, initial temperature t_{p0}, and final temperature t_{p1}.

Output: $\hat{g}$, the LUB on the optimal objective value of model (6.1) under $|\Omega|$ scenarios.

  1: Set $\hat{g}$ to a significantly large value.   ▷ Initialize $\hat{g}$

  2: Solve model (6.1) under m randomly chosen scenarios to obtain the current candidate solution X^* and let $\hat{X} \leftarrow X^*$.   ▷ Initialize $\hat{X}$

  3: **while** $\kappa \leq \kappa_{max}$ **do**

  4:       Solve model (6.2) under m randomly chosen scenarios to obtain X^*.

  5:       Solve model (6.1) under $|\Omega|$ scenarios with X fixed to X^* to obtain the optimal

  6:       objective value g^*.

  7:       **if** $g^* < \hat{g}$ **then**

  8:           $\hat{g} \leftarrow g^*$ ▷ Update the best-so-far objective value

  9:           $flag \leftarrow true$

10:       **else**   ▷ Accept the worse candidate solution with a probability

11:           Calculate the acceptance probability Pr according to Eq. (6.3).

12:           **if** $p_a \geq rand\,(0,1)$ **then**

13:               $flag \leftarrow true$

14:           **end if**

15:       **end if**

16:       **if** $flag = true$ **then**

17:           $\hat{X} \leftarrow X^*$ ▷ Accept the new candidate solution

18:       **end if**

19:       $t_p \leftarrow t_p \times \left(\frac{T_{p1}}{T_{p0}}\right)^{1/\kappa_{max}}$ ▷ Reduce temperature

20:       $\kappa \leftarrow \kappa + 1$

21: **end while**

In the hope of improving the solution quality, we integrate two search techniques into the iterative regularization scheme. The first one is *Simulated Annealing* (SA), a meta-heuristic technique for approximating the global optima of a cost function with multiple local optima by emulating annealing in metallurgy [201]. Physical annealing is the process of slowly cooling a heated solid to a state of minimum energy according to a particular

cooling schedule [202]. A key feature of the SA algorithm is that it can escape from local minima by allowing hill-climbing moves (moves to a worse solution) probabilistically according to the Metropolis criterion [203]. The acceptance probability is higher when the temperature t_p is higher, allowing the algorithm to explore more of the search space. The acceptance probability of a hill-climbing move is given by

$$Pr(\Delta E) = e^{-\frac{\Delta E}{t_p}}, \tag{6.3}$$

where ΔE represents the energy difference, i.e. the change in the cost function of the optimization problem under study. For a more details explanation of the SA algorithm, interested readers are referred to Refs. [202, 203].

The iterative regularization algorithm with a SA step in determining the acceptance of a new candidate solution is given by Algorithm 6. In this algorithm, we adopt the geometric cooling schedule where the temperature decreases by a constant factor [203].

We also consider another technique from meta-heuristics, *exploration and exploitation* [204], to find the approximate optimal objective value of the extensive formulation. Exploration means searching for unknown but promising solutions globally while exploitation ensures that the solution algorithm exploits the best-so-far solutions by searching in the vicinity of those solutions (local search). In the case of the iterative regularization, we use the original model (no regularization) during the exploration stage such that the solution algorithm can examine the solution space globally. The regularized model is used in the exploitation stage to find candidate optimal solutions around the promising solution obtained during the exploration stage. The iterative regularization scheme with the exploration and exploitation technique is described in Algorithm 7.

Algorithm 7: EE iterative regularization

Input: m, ρ, the maximum number of iterations κ_{max}, $|\Omega|$ failure scenarios, number of iterations for exploration κ_E.

Output: $\hat{g}$, the LUB on the optimal objective value of model (6.1) under $|\Omega|$ scenarios.

 1: Set $\hat{g}$ to a significantly large value.   ▷ Initialize $\hat{g}$

 2: **while** $\kappa \leq \kappa_{max}$ **do**

 3:     **if** $\kappa \leq \kappa_E$ **then**   ▷ Exploration stage: no regularization

 4:         Solve model (6.1) under m randomly chosen scenarios to obtain the current candidate

 5:         solution to the first stage problem $\boldsymbol{X}^*$.

 6:         Solve model (6.1) under $|\Omega|$ scenarios with $\boldsymbol{X}$ fixed to $\boldsymbol{X}^*$ to obtain the optimal

 7:         objective value g^*.

 8:     **else**   ▷ Exploitation stage: regularization is used

 9:         Solve model (6.2) under m scenarios with $\boldsymbol{X}$ fixed to $\hat{\boldsymbol{X}}$ to obtain $\boldsymbol{X}^*$.

10:         Solve model (6.1) under $|\Omega|$ scenarios with $\boldsymbol{X}$ fixed to $\boldsymbol{X}^*$ to obtain the optimal

11:         objective value g^*.

12:     **end if**

13:     **if** $g^* < \hat{g}$ **then**

14:         $\hat{g} \leftarrow g^*$ ▷ Update the best-so-far objective value

15:         $\hat{\boldsymbol{X}} \leftarrow \boldsymbol{X}^*$ ▷ Update the incumbent solution

16:     **end if**

17:     $\kappa \leftarrow \kappa + 1$

18: **end while**

6.4   Computational Experiments

6.4.1   Experimental Setup

The methodology is illustrated through computational experiments about synthesized 49-node interdependent power and gas networks. The computational experiments in this chapter are implemented in Julia/JuMP 0.21.2 [205, 206] on a Windows 10 desktop computer with a 3.40 GHz Intel Core i7-6700 CPU and 16.0 GB RAM. The models are solved with Gurobi 9.0.2 [207]. Relative MIPGap is set to 1.0×10^{-3}.

The time horizon $|\mathscr{T}|$ is set to 40 for the 49-node network. Each time period represents a 12-hour window. The number of repair crews is assumed to be 13 in the power network and 11 in the gas network. The hardening cost, restoration cost, and repair time of each component are estimated from various online data sources and literature with reasonable

assumptions.

Table 6.2: Cost and restoration time for different components. Units for cost and time are 1×10^5 USD and 4 hours, respectively. Power flow is measured by MW while gas flow is measured by MMCM/d. "∗" indicates that the value is adapted from the reference. "-" indicates assumed value.

Parameters	Value	Reference
Cost of hardening a demand node	0.70	[208]
Cost of repairing a demand node per time period	0.28	[209]
Cost of hardening a supply node	6.00	-
Cost of repairing a supply node per time period	1.50	-
Cost of hardening an arc	3.60	-
Cost of repairing an arc per time period	1.20	-
Power generation cost	27.78	[210]
Gas cost	3944.46	[211]
Penalty on unmet power load	277.80	-
Penalty on unmet gas demand	39444.60	-
Time required to repair a demand node	5.00	[212]∗
Time required to repair a supply node	8.00	-
Time required to repair an arc	6.00	-
Power to gas conversion factor	3.29	[213]
Gas to power conversion factor	172.47	[214]
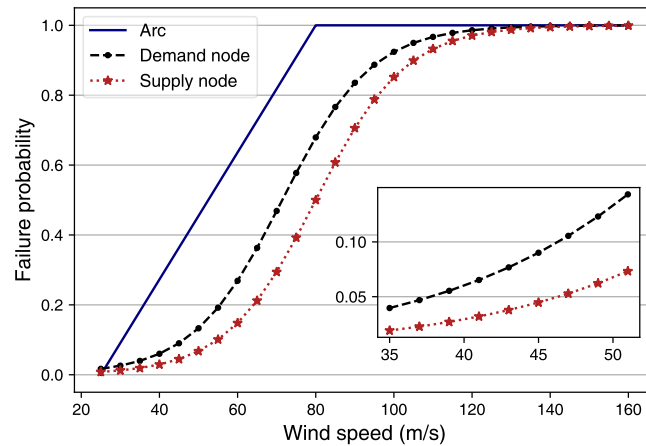
Figure 6.1: Fragility curve for different types of components

The disruptive event we consider for the numerical is hurricane. The wind fragility curves from Ref. [215] are used to determine the failure probability of a component affected by the hurricane (Fig. 6.1). Note that arcs in the gas network (gas pipelines) remain functional in all scenarios since the majority of pipelines are buried underground and are

not affected by the storms. Sustained wind speed is assumed to follow a Gaussian distribution with the mean equal to 44.8 m/s (approximately 100 mph), a typical value observed during a Category Two hurricane [216].

Due to the lack of information about the Euclidean distance between the components, we use the topological distance to determine the correlation between the failure probability of components according to Eq. (6.4):

$$\rho_{ij} = \begin{cases} 0.3, & d = 1 \\ 0.2, & d = 2 \\ 0.1, & d = 3 \\ 0, & d \geq 4 \end{cases} \tag{6.4}$$

6.4.2   Test Networks

The test network we use for the computational experiments is 49-node interdependent power and gas networks (Fig. 6.2) that consist of the IEEE RTS 24-bus test system [217] coupled with a 25-node gas network [6]. Gas-fired generators G1 to G4 depend on the supply from gas nodes J18, J13, J24, and J19 in gas network, respectively. The compressor at gas supply node J1 relies on the electricity from power node K22. The power flow is obtained by solving the DC optimal power flow model and the gas flow is calculated based on the diameter of pipelines and gas flow speed (100 m/s) [218]. The total demand in each network is assumed to be 65% of the total generation capacity. The demand and supply in gas network are scaled down 5 times so that the undamaged interdependent power-gas network has a feasible flow without unmet demand. For simplicity, we assume that the composition of the natural gas remains unchanged through the entire gas network so that the gas to power conversion factor is the same throughout the entire gas network. We also assume that the power generation cost is the same at all the generators.
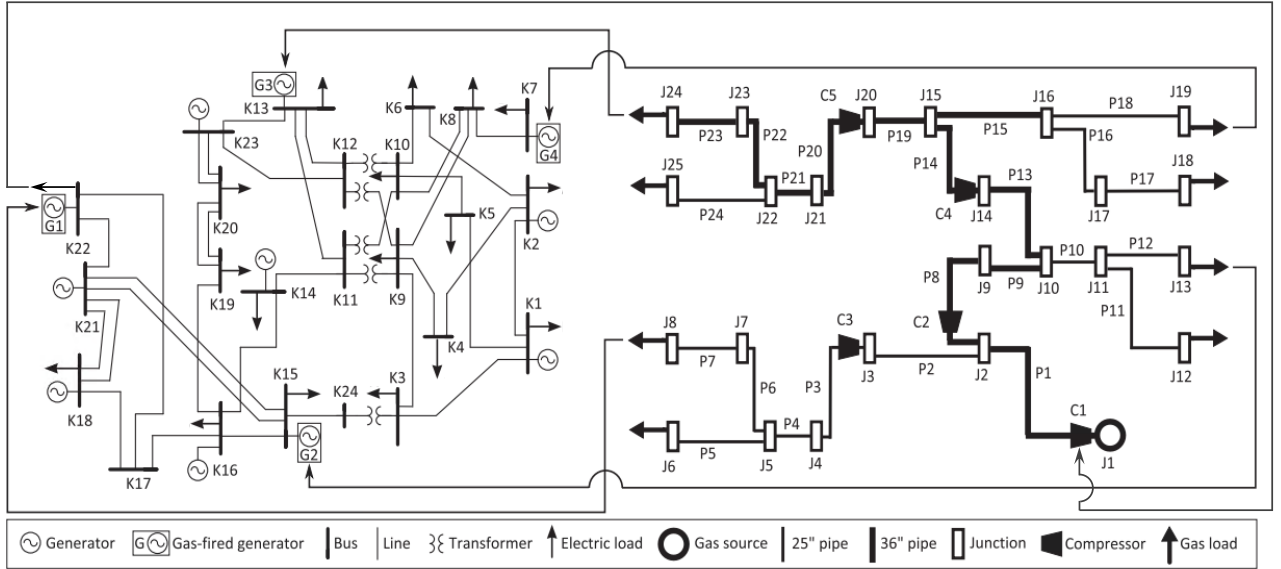
Figure 6.2: Schematic of 49-node interdependent power-gas networks (adapted from Ref. [6]). The power-gas networks are generated by coupling the IEEE RTS 24-bus test system (left) with a 25-node gas network (right) through gas-powered generators (G1 to G4) and an electric-powered compressor (C1).

6.4.3   Results

6.4.3.1   Performance of the Solution Techniques

We first examine the impact of the penalty strength ρ on the optimal objective value under different solution algorithms. The number of scenarios for model (6.1) of smaller scale m is 30 and the number of scenarios for the original model $|\Omega|$ is 3000. We run the three solution algorithms for $\kappa_{max} = 20$ times to obtain the optimal objective value. It can be observed in Fig. 6.3 that as ρ increases, the optimal objective obtained by iterative regularization based on EE descends stepwise towards a relatively wide valley and then ascends shapely. In comparison, the objective values obtained by iterative regularization based on SA and RW fluctuate, leading to several local minima in the space of ρ. The optimal objective values obtained by RW overlap with that by SA at multiple values of ρ, including the minimum values when $\rho \in [7 \times 10^4, 8 \times 10^4]$, but other than the overlapping points, SA consistently outperforms RW. Since the optimal objective values for the three
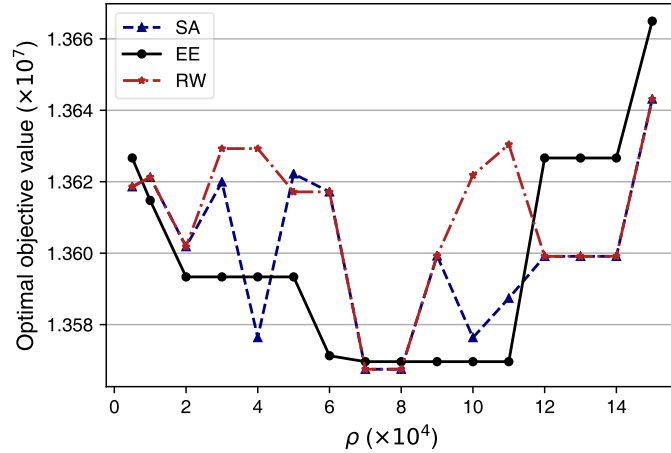
Figure 6.3: Trend of the optimal objective value over values of ρ for different algorithms

algorithms are all obtained at $\rho \in [7 \times 10^4, 8 \times 10^4]$, $\rho = 8 \times 10^4$ will be used when iterative regularization is employed in the rest of the computational experiments.

From the trend of the optimal objective value for the three algorithms depicted in Fig. 6.4, we can also see that the optimal objective values obtained by the three algorithms descend stepwise. The optimal objective values are achieved by iteration 15, so 15 iterations can be used. The trend of the optimal objective value obtained by iterative regularization based on SA and RW are almost identical except when the iteration number exceeds 15 at $\rho = 10 \times 10^4$ or $\rho = 11 \times 10^4$, thus SA-based algorithm is more likely to lead to better solutions given the same number of iterations. The EE-based iterative regularization algorithm descends to lower optimal objective value than SA- or RW-based iterative regularization. Also, due to the insensitivity of EE-based iterative regularization algorithm to the changes in ρ at the valley (Fig. 6.3), this algorithm is chosen to obtain the optimal objective value in the subsequent computational experiments where iterative regularization is employed.

The solution time for the three algorithms when $\rho = 8 \times 10^4$ are summarized in Table 6.3. We can observe that the solution times for the algorithms by the end of each iteration are very close to each other, with the total time for 20 iterations between 2.89 to
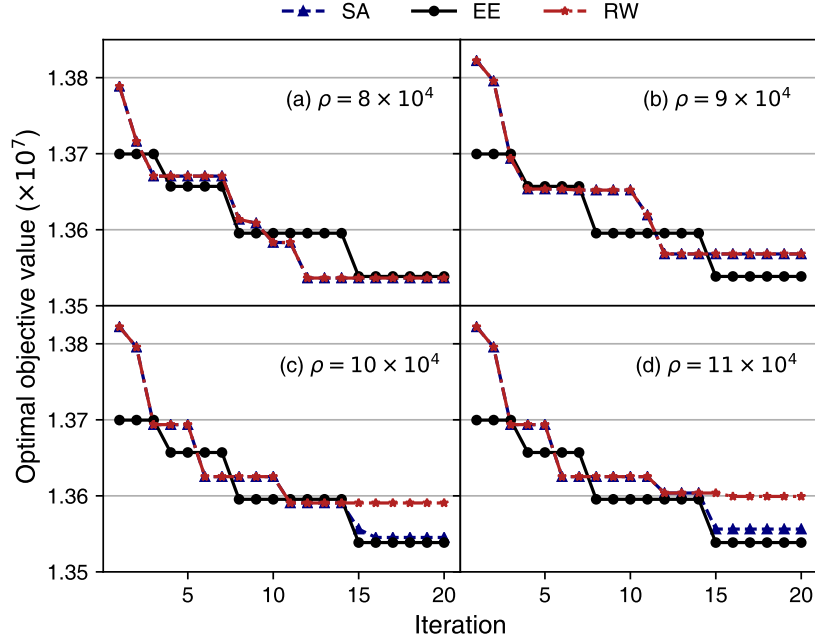
Figure 6.4: Trend of the optimal objective value over iterations for different algorithms

2.94 hours and the average time for one iteration of approximately 0.15 hour. The solution time for each algorithm when ρ takes other values is omitted here because the solution time varies little with the value of ρ.

Table 6.3: Solution time for each algorithm

Iter.	Time (hrs)		
	SA	EE	RW
1	0.16	0.15	0.15
4	0.59	0.58	0.58
8	1.17	1.16	1.18
12	1.76	1.74	1.77
16	2.33	2.32	2.34
20	2.90	2.89	2.94

6.4.3.2 Network Hardening Budget

We next explore the impact of the budget for network hardening B on the optimal objective value $\hat{g}$ of the IPRIN problem. The optimal objective values are obtained by iterative

113

regularization with EE where the number of scenarios for model (6.1) of smaller scale m is 30, the number of scenarios for original model $|\Omega|$ is 3000, and the number of iterations κ_{max} is 20. The optimal objective value $\hat{g}$ for different values of budget B is depicted in Fig. 6.5. It can be observed that as the budget increases from zero, the optimal objective value first decreases rapidly. When the budget grows over around $\$1.2 \times 10^7$, the decrease in the optimal objective value slows down until it eventually levels off from $B = 3.0 \times 10^7$ onwards since the budget is large enough to harden all the components. It is worthwhile to note that when the budget is lower than approximately $\$1.2 \times 10^7$, the relationship between the optimal objective value and the budget is approximately linear, indicating that every unit increase in the budget for hardening will lead to 0.7761 unit decrease in the total cost at the restoration stage.
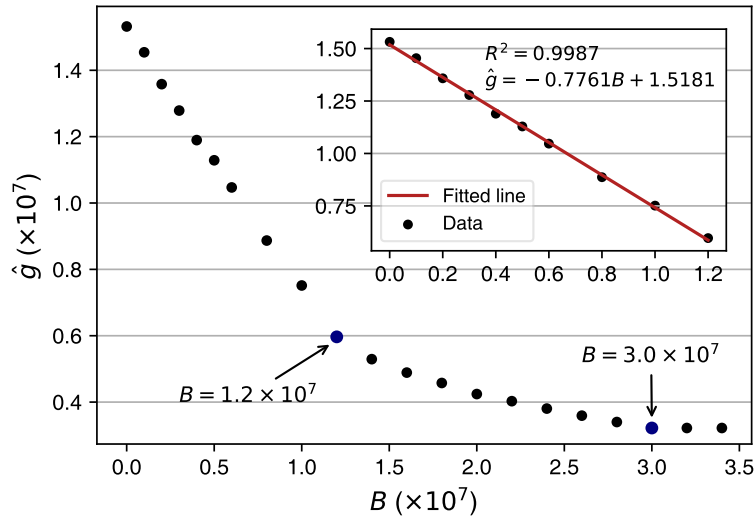
Figure 6.5: Optimal objective value $\hat{g}$ for different budgets for network hardening budget B

6.4.3.3 Expected Value of Perfect Information

To evaluate the benefits of the proposed approach, we compare the solution to the stochastic program (SP), i.e. the *here-and-now* solution, to the *wait-and-see* solution (WS). WS is obtained by solving the deterministic problem under each scenario independently as if decision-makers can wait and see the realization of the disruptive event. The difference

between SP and WS is referred to as the *expected value of perfect information* (EVPI), which measures the value of complete (and accurate) forecast for the future [219, 40]. Formally, WS is calculated by [40]

$$\text{WS} = \mathbb{E}_{\omega} \left[\min_{\boldsymbol{X}} g\left(\boldsymbol{X}, \omega\right) \right] \tag{6.5a}$$

$$= \mathbb{E}_{\omega} \left[g\left(\bar{\boldsymbol{X}}\left(\omega\right), \omega\right) \right] \tag{6.5b}$$

where $g\left(\cdot\right)$ represents the objective function and $\bar{\boldsymbol{X}}\left(\omega\right)$ represents the optimal solution to the first stage decisions after the realization of a particular scenario. In this analysis, we use $|\Omega| = 3000$ scenarios to estimate the value of WS and SP. SP is estimated as the LUB among the 20 upper bounds obtained by feeding the solution to model (6.1) under 200 scenarios to model (6.1) under 3000 scenarios. The EPVI is found to be

$$\text{EVPI} = \text{SP} - \text{WS} = 13,571,256 - 10,395,870 = 3,175,386 \tag{6.6}$$

The EVPI is approximately 23.4% of the SP. This implies that a decision maker would be ready to pay $3,078,544 to collect perfect information about the damage of the disruptive event to the interdependent power and gas networks.

6.5   Summary

Accounting for the uncertainty around human-induced or natural disastrous events is important to achieve the optimal protection and restoration planning of infrastructure systems. This chapter proposes a two-stage mixed-integer stochastic programming approach to solve the problem of combining disaster preparation and restoration decisions to enhance the protection and restoration planning of IISs. Novel solution algorithms that integrate random walk, SA, exploration and exploitation techniques and iterative regularization are devised to achieve a near-optimal solution within a reasonable amount of time for rela-

tively large-scale IISs. Results show that the EE-based iterative regularization outperforms the other two algorithms in finding the LUB. Managerial insights for the interdependent infrastructure networks are presented, including the the expected value of the perfect information and impact of the budget for network hardening on the optimal objective value.

Chapter 7

Summary and Future Work

7.1   Summary of Contributions

The overall objective of this dissertation is to develop and apply Bayesian models, stochastic network modeling and optimization models to assess and improve the resilience of complex infrastructure systems, particularly interdependent systems. The various contributions made in this dissertation are summarized below.

Hierarchical Bayesian model for the serviceability assessment of infrastructure systems under uncertainty. First, this dissertation develops a method based on HBM to address the challenge due to data scarcity and leverages potential availability of data in the future to provide an updated probabilistic evaluation of infrastructure performance. Second, this dissertation proposes an approach where both epistemic uncertainty and aleatory uncertainty about component failure probability are modeled. In the proposed method, the epistemic uncertainty is modeled with a distribution whose parameters are also described by a distribution. The Bayesian updating approach is integrated into the Monte Carol simulations for evaluating the serviceability of infrastructure networks.

Hierarchical Bayesian kernel regression model. First, a novel model, hierarchical Bayesian kernel model (HBKM), is proposed. This model can 1) improve the estimation accuracy of parameters under small data due to 1) its ability to "borrow strength" across different but similar groups of datasets, and 2) reduce the subjectivity of assigning a point estimate to prior parameters by replacing the point estimate with a hyperprior that is updated using the entire dataset. As a Bayesian approach, this model can provide a probabilistic estimate of the parameters of interest, which is highly useful for decision making under certainty or probabilistic risk analysis. Second, this dissertation validates the HBKM on the historical power outage data and multiple synthetic grouped count data against compet-

117

itive models, including generalized linear model, Gaussian process regression, and kernel regression. HBKM can outperform other models on some of the simulated datasets depending on 1) the relationship between the response variables and the predictors and 2) the number of data points in each group and the number of groups.

Scalable inference of hierarchical Bayesian regression model for grouped count data. First, a hierarchical Bayesian regression model for grouped count data is proposed. The proposed model is learned using a new scalable approximate Gibbs sampler that targets relatively high-dimensional posterior distributions. The data likelihood is approximated with a Gaussian distribution to obtain closed-form full conditional posterior distribution of coefficients. To improve the convergence quality of posterior samples of the coefficients, i.e. reduce the approximation error as much as possible, we derive the exact mean and variance of the approximate Gaussian distribution rather than using the approximate mean and variance. Second, we validate the approximate Gibbs samplers against the state-of-the-art sampler, the No-U-Turn-Sampler (NUTS), on multiple grouped count datasets. The approximate Gibbs sampler can achieve slightly higher inference accuracy and faster speed than NUTS, particularly for large datasets.

Modeling interdependencies using the stochastic block model. The first contribution is that a modified stochastic block model is proposed to evaluate uncertain network interdependencies. Specifically, the use of stochastic block models provides a probabilistic characterization of interdependency links between infrastructure networks. Second, this dissertation proposes a formulation for estimating the likelihood that an interdependency link exists using multiple predictors to represent major factors influencing the presence of interdependency links. To avoid zero values in normalizing the variables that will be used in the denominator, a variant of min-max normalization called truncated min-max normalization is developed.

Two-stage stochastic program for the protection and restoration of IISs. First, a two-stage stochastic mixed-integer program is developed to optimize the decisions on the

integrated protection and restoration of interdependent critical infrastructure systems under uncertainty around the disruptive events. To the best of our knowledge, our study is the first effort in the literature to formulate this integrated protection and restoration of interdependent infrastructure networks under uncertainty. Second, We solve the proposed two-stage mixed-integer program for relatively large interdependent networks by integrating iterative regularization into the classic sampling average approach. The original model with a relatively small number of scenarios is first solved to obtain a plausible optimal solution. The quality of the plausible optimal solution is improved by iteratively solving the regularized model with an increasing number of scenarios within the allowed solution time. Thus, this method is able to provide a sufficiently high-quality solution to the integrated model within a reasonable amount of time. Third, the numerical analysis based on realistic case studies demonstrates the value of the stochastic model and the effectiveness of the proposed solution method. The proposed model can serve as a useful tool to inform the protection decisions of IISs.

7.2   Future Work

The following directions for future research are identified based on the work presented in this dissertation.

Hierarchical Bayesian Model for the Serviceability Assessment of Infrastructure Systems. Future directions of this research effort can focus on refining the model for estimating the failure probability of components under the flood hazard using new data, which can be collected from repair records or through the use of sensing technologies that provide real-time detection of pipe failures or leakages [220]. Another direction for further research is to incorporate interdependencies among infrastructure systems into the Bayesian updating framework for assessing the serviceability because WDSs are often coupled with the power grids [8]. For example, flooding caused by water pipe breaks can damage closely located power distribution facilities, which may cause the pumping stations to fail due to

loss of power supply. As such, the interdependency-related failure should be included to provide a comprehensive assessment of the serviceability of water distribution systems.

Hierarchical Bayesian Kernel Model. The proposed HBKM model can be a useful tool for utilities and disaster responders to understand the response mechanism of a certain infrastructure network and plan accordingly. As the ultimate goal of this part of the dissertation is to model the recovery process of infrastructure networks and communities as a function of different infrastructure and social characteristics, future work can incorporate additional information such as demographic data to evaluate the impact of social vulnerability on the recovery process and assess the performance of HBKMs in multiple regression where a set of covariates will be utilized to build the models.

Scalable inference of hierarchical Bayesian models. As the proposed inference algorithm based on approximate Gibbs is not applicable to datasets with zero counts, future work will explore scalable inference in hierarchical Bayesian models for zeros-inflated count data.

Modeling Interdependencies Using Probabilistic Network Models. A limitation of modeling interdependency using the adapted stochastic block model is that topology-based metrics for resilience provide a basic understanding of the dynamics of ICIs after disruptive events. Future work can consider building a full network flow model for a more accurate representation of the performance and behavior of ICI after a disruption under uncertainty. Additionally, incorporating more nodal attributes in the probabilistic model for estimating interdependency links will help with the interpretation of the model.

Stochastic Programming Approach for the Protection and Restoration of Infrastructure Networks. There are several possible extensions of this work. First, the stochastic formulation can be improved by incorporating the uncertainty around repair times into the model. Future work can also consider small-scale network reconfiguration, e.g. installing temporary power lines in the power system, to accelerate the restoration. Furthermore, as additional disruptions may occur before the infrastructure systems are fully restored, e.g. a

second winter storm can strike infrastructure networks not long after they are damaged by the first, the current formulation can be extended into a multi-stage stochastic program to solve the corresponding decision making problem under uncertainty.

Due to the complexity of the study required for each objective in this dissertation, the tasks of estimation and the task of optimization for enhancing the resilience of infrastructure systems are performed separately. This decoupled approach is the standard paradigm in many analytics problems, in which the estimation or prediction of unknown parameters from machine learning models are not appropriately integrated into the downstream optimization problems [221]. However, one can adopt the smart "Predict, then Optimize" (SPO) approach in which the machine learning models are trained to minimize the decision error rather than the prediction/estimation error [221]. By following the SPO paradigm, long-term research can investigate integrating Bayesian methods or other machine learning methods for estimating recovery state into a multi-stage stochastic program or Markov decision process for optimizing decisions about preparedness and restoration, so that the resilience of infrastructure systems under uncertainty can be further enhanced.

BIBLIOGRAPHY

[1] Devanandham Henry and Jose Emmanuel Ramirez-Marquez. Generic metrics and quantitative approaches for system resilience as a function of time. *Reliability Engineering & System Safety*, 99:114–122, 2012.

[2] Xiaoge Zhang, Sankaran Mahadevan, Shankar Sankararaman, and Kai Goebel. Resilience-based network design under uncertainty. *Reliability Engineering & System Safety*, 169:364–379, 2018.

[3] Yu Wang, Jin-Zhu Yu, and Hiba Baroud. Generating synthetic critical interdependent infrastructure networks. *Submitted manuscript*, 2020.

[4] FEMA. Multi-hazard loss estimation methodology: earthquake model. *Washington, DC, USA: Federal Emergency Management Agency*, 2013.

[5] Federal Emergency Management Agency Mitigation Division. Hazus-MH technical manual: flood model. Technical report, Department of Homeland Security, Washington D.C., 2013. URL https://www.fema.gov/media-library-data/20130726-1820-25045-8292/hzmh2_1_fl_tm.pdf.

[6] Anatoly Zlotnik, Line Roald, Scott Backhaus, Michael Chertkov, and Göran Andersson. Coordinated scheduling for interdependent electric power and natural gas infrastructures. *IEEE Transactions on Power Systems*, 32(1):600–610, 2016.

[7] Irene Eusgeld, Wolfgang Kröger, Giovanni Sansavini, Markus Schläpfer, and Enrico Zio. The role of network theory and object-oriented modeling within a framework for the vulnerability analysis of critical infrastructures. *Reliability Engineering & System Safety*, 94(5):954–963, 2009.

[8] Min Ouyang. Review on modeling and simulation of interdependent critical infrastructure systems. *Reliability engineering & System safety*, 121:43–60, 2014.

[9] Min Ouyang, Min Xu, Chi Zhang, and Shitong Huang. Mitigating electric power system vulnerability to worst-case spatially localized attacks. *Reliability Engineering & System Safety*, 165:144–154, 2017.

[10] ASCE. Report card for America's infrastructure. Technical report, 2017. URL https://www.infrastructurereportcard.org.

[11] Alessandro Vespignani. Complex networks: The fragility of interdependency. *Nature*, 464(7291):984, 2010.

[12] Amir Bashan, Yehiel Berezin, Sergey V Buldyrev, and Shlomo Havlin. The extreme vulnerability of interdependent spatially embedded networks. *Nature Physics*, 9(10): 667, 2013.

[13] Michael M Danziger, Louis M Shekhtman, Amir Bashan, Yehiel Berezin, and Shlomo Havlin. Vulnerability of interdependent networks and networks of networks. In *Interconnected Networks*, pages 79–99. Springer, 2016.

[14] White House. Presidential policy directive: Critical infrastructure security and resilience. *Washington D.C.*, 2013.

[15] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009.

[16] Zhenyu Yan and Yacov Y Haimes. Cross-classified hierarchical Bayesian models for risk-based analysis of complex systems under sparse data. *Reliability Engineering & System Safety*, 95(7):764–776, 2010.

[17] Jin-Zhu Yu and Hiba Baroud. Quantifying community resilience using hierarchical Bayesian kernel methods: A case study on recovery from power outages. *Risk Analysis*, 39:1930–1948, 2019.

[18] Joshua Mullins, You Ling, Sankaran Mahadevan, Lin Sun, and Alejandro Strachan. Separation of aleatory and epistemic uncertainty in probabilistic model validation. *Reliability Engineering & System Safety*, 147:49–59, 2016.

[19] Sarah G Nurre, Burak Cavdaroglu, John E Mitchell, Thomas C Sharkey, and William A Wallace. Restoring infrastructure systems: An integrated network design and scheduling (inds) problem. *European Journal of Operational Research*, 223(3):794–806, 2012.

[20] Burak Cavdaroglu, Erik Hammel, John E Mitchell, Thomas C Sharkey, and William A Wallace. Integrating restoration and scheduling decisions for disrupted interdependent infrastructure systems. *Annals of Operations Research*, 203(1):279–294, 2013.

[21] Andrés D González, Leonardo Dueñas-Osorio, Mauricio Sánchez-Silva, and Andrés L Medaglia. The interdependent network design problem for optimal infrastructure system restoration. *Computer-Aided Civil and Infrastructure Engineering*, 31(5):334–350, 2016.

[22] Nafiseh Ghorbani-Renani, Andrés D González, Kash Barker, and Nazanin Morshedlou. Protection-interdiction-restoration: Tri-level optimization for enhancing interdependent network resilience. *Reliability Engineering & System Safety*, page 106907, 2020.

[23] Crawford S Holling. Resilience and stability of ecological systems. *Annual review of ecology and systematics*, 4(1):1–23, 1973.

[24] Bilal M Ayyub. Practical resilience metrics for planning, design, and decision making. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 1(3):04015008, 2015.

[25] National Infrastructure Advisory Council. *Critical infrastructure resilience: Final report and recommendations*. National Infrastructure Advisory Council (US), 2009.

[26] National Research Council. *Disaster resilience: A national imperative*. National Academies Press, Washington, DC, 2012.

[27] Presidential Policy Directive. *Critical infrastructure security and resilience*. White House, 2013.

[28] Yacov Y Haimes. On the definition of resilience in systems. *Risk Analysis: An International Journal*, 29(4):498–501, 2009.

[29] Elyssa L Mooney, Yasser Almoghathawi, and Kash Barker. Facility location for recovering systems of interdependent networks. *IEEE Systems Journal*, 13(1):489–499, 2018.

[30] Yasser Almoghathawi, Kash Barker, and Laura A Albert. Resilience-driven restoration model for interdependent infrastructure networks. *Reliability Engineering & System Safety*, 185:12–23, 2019.

[31] Xiaoge Zhang. *Machine Learning and Optimization Models to Assess and Enhance System Resilience*. PhD thesis, Vanderbilt University, 2019.

[32] Carolyn J Anderson, Stanley Wasserman, and Katherine Faust. Building stochastic blockmodels. *Social networks*, 14(1-2):137–161, 1992.

[33] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[34] Aleš Žiberna. Blockmodeling of multilevel networks. *Social networks*, 39:46–61, 2014.

[35] Pierre Barbillon, Sophie Donnet, Emmanuel Lazega, and Avner Bar-Hen. Stochastic block models for multiplex networks: an application to a multilevel network of researchers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1):295–314, 2017.

[36] Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

[37] Kevin Xu. Stochastic block transition models for dynamic networks. In *Artificial Intelligence and Statistics*, pages 1079–1087, 2015.

[38] Eric D Kolaczyk. *Topics at the Frontier of Statistics and Network Analysis:(re) visiting the Foundations*. Cambridge University Press, 2017.

[39] Michael W Mahoney. Lecture notes in spectral graph methods, 2015.

[40] John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[41] Alexander Shapiro and Andy Philpott. A tutorial on stochastic programming. *Manuscript. Available at www2. isye. gatech. edu/ashapiro/publications. html*, 17, 2007.

[42] Simge Küçükyavuz and Suvrajeet Sen. An introduction to two-stage stochastic mixed-integer programming. In *Leading Developments from INFORMS Communities*, pages 1–27. INFORMS, 2017.

[43] Stephen Boyd, Lin Xiao, Almir Mutapcic, and Jacob Mattingley. Notes on decomposition methods. *Notes for EE364B, Stanford University*, pages 1–36, 2007.

[44] Michael Betancourt and Mark Girolami. Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79:30, 2015.

[45] Department of Homeland Security. National infrastructure protection plan: Partnering to enhance protection and resiliency. *Washington D.C.*, 2009.

[46] D Aldred. Urbanization: a major driver of infrastructure spending. industry trends: Infrastructure. *City Perspectives*, 2012.

[47] Min Ouyang. Critical location identification and vulnerability analysis of interdependent infrastructure systems under spatially localized attacks. *Reliability Engineering & System Safety*, 154:106–116, 2016.

[48] Daniele Laucelli and Orazio Giustolisi. Vulnerability assessment of water distribution networks under seismic actions. *Journal of Water Resources Planning and Management*, 141(6):04014082, 2014.

[49] Sybil Derrible. An approach to designing sustainable urban infrastructure. *MRS Energy & Sustainability*, 5, 2018.

[50] ASCE. 2017 infrastructure report: wastewater. Technical report, American Society of Civil Engineers ASCE, 2017. URL https://www.infrastructurereportcard.org/cat-item/wastewater/.

[51] ASCE. 2017 infrastructure report: drinking water. Technical report, American Society of Civil Engineers ASCE, 2017. URL https://www.infrastructurereportcard.org/cat-item/drinking-water/.

[52] Keith Miller, Kristina Costa, and Donna Cooper. How to upgrade and maintain our nation's wastewater and drinking water infrastructure. Technical report, Center for American Progress, 2012. URL https://www.americanprogress.org/wp-content/uploads/2012/10/MillerWaterInfrastructureReport.pdf.

[53] Wisinee Wisetjindawat, Amirhassan Kermanshah, Sybil Derrible, and Motohiro Fujita. Stochastic modeling of road system performance during multihazard events: flash floods and earthquakes. *Journal of infrastructure systems*, 23(4):04017031, 2017.

[54] Shangjia Dong, Tianbo Yu, Hamed Farahmand, and Ali Mostafavi. Bayesian modeling of flood control networks for failure cascade characterization and vulnerability assessment. *Computer-Aided Civil and Infrastructure Engineering*, 2019.

[55] Golam Kabir, Solomon Tesfamariam, and Rehan Sadiq. Predicting water main failures using Bayesian model averaging and survival modelling approach. *Reliability Engineering & System Safety*, 142:498–514, 2015.

[56] Takao Adachi and Bruce R Ellingwood. Serviceability assessment of a municipal water system under spatially correlated seismic intensities. *Computer-Aided Civil and Infrastructure Engineering*, 24(4):237–248, 2009.

[57] Yu Wang and Thomas D O'Rourke. Seismic performance evaluation of water supply systems. 2008.

[58] Qing Shuang, Mingyuan Zhang, and Yongbo Yuan. Node vulnerability of water distribution networks under cascading failures. *Reliability Engineering & System Safety*, 124:132–141, 2014.

[59] National Weather Service. Summary of natural hazard statistics for 2016 in the United States. Technical report, National Oceanic and Atmospheric Administration, 2017. URL http://www.nws.noaa.gov/os/hazstats/sum16.pdf.

[60] A Kermanshah, S Derrible, and M Berkelhammer. Using climate models to estimate urban vulnerability to flash floods. *Journal of Applied Meteorology and Climatology*, 56(9):2637–2650, 2017.

[61] Miroslav Nastev and Nikolay Todorov. Hazus: A standardized methodology for flood risk assessment in Canada. *Canadian Water Resources Journal*, 38(3):223–231, 2013.

[62] Jonathan WF Remo, Nicholas Pinter, and Moe Mahgoub. Assessing illinois's flood vulnerability using Hazus-MH. *Natural Hazards*, 81(1):265–287, 2016.

[63] Claudia Copeland. Hurricane-damaged drinking water and wastewater facilities: impacts, needs, and response. Congressional Research Service, Library of Congress, 2005.

[64] Shridhar Yamijala, Seth D Guikema, and Kelly Brumbelow. Statistical models for the analysis of water distribution system pipe break data. *Reliability Engineering & System Safety*, 94(2):282–293, 2009.

[65] Jin-Zhu Yu and Hiba Baroud. Quantifying community resilience using hierarchical Bayesian kernel methods: A case study on recovery from power outages. *Risk Analysis*, 39(9):1930–1948, 2019.

[66] Weiwen Peng, Hong-Zhong Huang, Min Xie, Yuanjian Yang, and Yu Liu. A Bayesian approach for system reliability analysis with multilevel pass-fail, lifetime and degradation data sets. *IEEE Transactions on Reliability*, 62(3):689–699, 2013.

[67] Iris Tien and Armen Der Kiureghian. Algorithms for Bayesian network modeling and reliability assessment of infrastructure systems. *Reliability Engineering & System Safety*, 156:134–147, 2016.

[68] Nita Yodo, Pingfeng Wang, and Zhi Zhou. Predictive resilience analysis of complex systems using dynamic Bayesian networks. *IEEE Transactions on Reliability*, 66 (3):761–770, 2017.

[69] Hiba Baroud and Kash Barker. A Bayesian kernel approach to modeling resilience-based network component importance. *Reliability Engineering & System Safety*, 170:10–19, 2018.

[70] HA Jensen and DJ Jerez. A Bayesian model updating approach for detection-related problems in water distribution networks. *Reliability Engineering & System Safety*, 185:100–112, 2019.

[71] Nathan O Siu and Dana L Kelly. Bayesian parameter estimation in probabilistic risk assessment. *Reliability Engineering & System Safety*, 62(1-2):89–116, 1998.

[72] Royce A Francis, Seth D Guikema, and Lucas Henneman. Bayesian belief networks for predicting drinking water distribution system pipe breaks. *Reliability Engineering & System Safety*, 130:1–11, 2014.

[73] Alessandro Pagano, Raffaele Giordano, Ivan Portoghese, Umberto Fratino, and Michele Vurro. A Bayesian vulnerability assessment tool for drinking water mains under extreme events. *Natural hazards*, 74(3):2193–2227, 2014.

[74] Golam Kabir, Solomon Tesfamariam, Alex Francisque, and Rehan Sadiq. Evaluating risk of water mains failure using a bayesian belief network model. *European Journal of Operational Research*, 240(1):220–234, 2015.

[75] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[76] Valen E Johnson, Ann Moosman, and Paul Cotter. A hierarchical model for estimating the early reliability of complex systems. *IEEE Transactions on Reliability*, 54 (2):224–231, 2005.

[77] Ming Yang, Faisal Khan, Leonard Lye, and Paul Amyotte. Risk assessment of rare events. *Process Safety and Environmental Protection*, 98:102–108, 2015.

[78] António Ramos Andrade and Paulo F Teixeira. Statistical modelling of railway track geometry degradation using hierarchical bayesian models. *Reliability Engineering & System Safety*, 142:169–183, 2015.

[79] Hongyang Yu, Faisal Khan, and Brian Veitch. A flexible hierarchical Bayesian modeling technique for risk analysis of major accidents. *Risk analysis*, 37(9):1668–1682, 2017.

[80] American Lifelines Alliance. Seismic Fragility Formulations for Water Systems Part 1 - Guideline. Technical report, FEMA, ASCE, Washington D.C., 2001. URL https://www.americanlifelinesalliance.com/pdf/Part_1_Guideline.pdf.

[81] Russell Blessing, Antonia Sebastian, and Samuel D Brody. Flood risk delineation in the united states: How much loss are we capturing? *Natural Hazards Review*, 18 (3):04017002, 2017.

[82] Gail M Atkinson and David M Boore. Recent trends in ground motion and spectral response relations for North America. *Earthquake Spectra*, 6(1):15–35, 1995.

[83] James L Beck and Siu-Kui Au. Bayesian updating of structural models and reliability using Markov chain Monte Carlo simulation. *Journal of Engineering Mechanics*, 128(4):380–391, 2002.

[84] Sai Hung Cheung and James L Beck. Bayesian model updating using hybrid monte carlo simulation with application to structural dynamic models with many uncertain parameters. *Journal of engineering mechanics*, 135(4):243–255, 2009.

[85] Dana L Kelly and Curtis L Smith. Bayesian inference in probabilistic risk assessment—the current state of the art. *Reliability Engineering & System Safety*, 94(2): 628–643, 2009.

[86] Stan Development Team and Others. Rstan: the R interface to Stan. *R package version*, 2(1), 2016.

[87] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[88] Nathan P Lemoine. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, 128(7):912–928, 2019.

[89] Min Wang and Tsuyoshi Takada. Macrospatial correlation model of seismic ground motions. *Earthquake Spectra*, 21(4):1137–1156, 2005.

[90] Sagy Cohen, G Robert Brakenridge, Albert Kettner, Bradford Bates, Jonathan Nelson, Richard McDonald, Yu-Fen Huang, Dinuke Munasinghe, and Jiaqi Zhang. Estimating floodwater depths from flood inundation maps and topography. *JAWRA Journal of the American Water Resources Association*, 2017.

[91] Federal Emergency Management Agency Mitigation Division. *Flood map service*. Department of Homeland Security, Washington D.C., 2018. URL https://msc.fema.gov.

[92] National Oceanic and Atmospheric Administration. *Advanced hydrologic prediction service flood inundation mapping*. Department of Commerce, Washington D.C., 2018. URL https://water.weather.gov/ahps/inundation.php.

[93] Eric Tate, Cristina Muñoz, and Jared Suchan. Uncertainty and sensitivity analysis of the HAZUS-MH flood model. *Natural Hazards Review*, 16(3):04014030, 2014.

[94] Stefan Reese and Doug Ramsay. Riskscape: flood fragility methodology. *Wellington, New Zealand. National Institute of Water and Atmospheric Research*, page 42, 2010.

[95] Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6): 345, 1962.

[96] Stephen Warshall. A theorem on boolean matrices. *Journal of the ACM (JACM)*, 9 (1):11–12, 1962.

[97] Howard HM Hwang, Huijie Lin, and Masanobu Shinozuka. Seismic performance assessment of water delivery systems. *Journal of Infrastructure Systems*, 4(3):118–125, 1998.

[98] SC Harmsen, AD Frankel, and MD Petersen. Deaggregation of US seismic hazard sources: The 2002 update. open file report 2003-03-440. *US Geological Survey*, 2003.

[99] National Oceanic and Atmospheric Administration. *PDS-based precipitation frequency estimates with 90% confidence intervals (in inches)*. Department of Commerce, Washington D.C., 2018. URL https://hdsc.nws.noaa.gov/hdsc/pfds/pfds_map_cont.html.

[100] Stan Development Team and Others. Stan modeling language users guide and reference manual. *Technical report*, 2016.

[101] Xiaoge Zhang, Andrew Adamatzky, Hai Yang, Sankaran Mahadaven, Xin-She Yang, Qing Wang, and Yong Deng. A bio-inspired algorithm for identification of critical components in the transportation networks. *Applied Mathematics and Computation*, 248:18–27, 2014.

[102] Yi-Ping Fang, Nicola Pedroni, and Enrico Zio. Resilience-based component importance measures for critical infrastructure network systems. *IEEE Transactions on Reliability*, 65(2):502–512, 2016.

[103] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[104] RX Zhong, KY Fu, A Sumalee, D Ngoduy, and WHK Lam. A cross-entropy method and probabilistic sensitivity analysis framework for calibrating microscopic traffic models. *Transportation Research Part C: Emerging Technologies*, 63:147–169, 2016.

[105] Stephanie E Chang. Urban disaster recovery: a measurement framework and its application to the 1995 kobe earthquake. *Disasters*, 34(2):303–327, 2010.

[106] J Liu, Z Shi, D Lu, and Y Wang. Measuring and characterizing community recovery to earthquake: The case of 2008 wenchuan earthquake, china. *Natural Hazards and Earth System Sciences, Discussion paper*, 2017.

[107] Greg M Allenby, Peter E Rossi, and RE McCulloch. Hierarchical Bayes models: A practitioners guide. grover r, vriens m, eds. *SSRN Electron J*, 2005.

[108] William J Browne, David Draper, et al. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian analysis*, 1(3):473–514, 2006.

[109] Cameron A MacKenzie, Theodore B Trafalis, and Kash Barker. A Bayesian beta kernel model for binary classification and online learning problems. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(6):434–449, 2014.

[110] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[111] Luis Montesano and Manuel Lopes. Learning grasping affordances from local visual descriptors. In *2009 IEEE 8th international conference on development and learning*, pages 1–6. IEEE, 2009.

[112] Hiba Baroud, Royce Francis, and Kash Barker. Data-driven methods for the risk analysis of global supply chains. In *13th International Conference on Probabilistic Safety Assessment and Management (PSAM 13)*, 2016.

[113] Hiba Baroud, Kash Barker, and Raychal Lurvey. Bayesian kernel model for disruptive event data. In *IIE Annual Conference. Proceedings*, page 1777. Institute of Industrial and Systems Engineers (IISE), 2013.

[114] Hiba Baroud. Bayesian kernel methods for the risk analysis and resilience modeling of critical infrastructure systems. 2015.

[115] Milton Friedman and Leonard J Savage. The utility analysis of choices involving risk. *Journal of political Economy*, 56(4):279–304, 1948.

[116] Yu Marco Nie, Xing Wu, and Tito Homem-de Mello. Optimal path problems with second-order stochastic dominance constraints. *Networks and Spatial Economics*, 12(4):561–587, 2012.

[117] Steven M Quiring, Laiyin Zhu, and Seth D Guikema. Importance of soil and elevation characteristics for modeling hurricane-induced power outages. *Natural hazards*, 58(1):365–390, 2011.

[118] Marcelo Figueroa-Candia, Frank A Felder, and David W Coit. Resiliency-based optimization of restoration policies for electric power distribution systems. *Electric Power Systems Research*, 161:188–198, 2018.

[119] Ali Arab, Amin Khodaei, Zhu Han, and Suresh K Khator. Proactive recovery of electric power assets for resiliency enhancement. *Ieee Access*, 3:99–109, 2015.

[120] Emily Shapiro and Julia Jacobo. 8 dead after Irma knocks out air conditioning at Florida nursing home. *American Broadcasting Company*. URL https://abcnews.go.com/US/dead-florida-nursing-home-irma-tore-state/story?id=49817477.

[121] Jin-Zhu Yu, Mackenzie Whitman, Amirhassan Kermanshah, and Hiba Baroud. A hierarchical Bayesian approach for assessing infrastructure networks serviceability under uncertainty: A case study of water distribution systems subject to earthquake and flood risk. *Submitted manuscript*, 2019.

[122] Yang Ni, Francesco C Stingo, Min Jin Ha, Rehan Akbani, and Veerabhadran Baladandayuthapani. Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics. *Journal of the American Statistical Association*, 114(525):48–60, 2019.

[123] Andrew Gelman. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435, 2006.

[124] Peter Congdon. *Bayesian statistical modelling*, volume 704. John Wiley & Sons, 2007.

[125] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[126] James A Fordyce, Zachariah Gompert, Matthew L Forister, and Chris C Nice. A hierarchical Bayesian approach to ecological count data: a flexible tool for ecologists. *PloS one*, 6(11):e26785, 2011.

[127] Ritabrata Dutta, Paul Blomstedt, and Samuel Kaski. Bayesian inference in hierarchical models by combining independent posteriors. *arXiv preprint arXiv:1603.09272*, 2016.

[128] Khalifeh AlJadda, Mohammed Korayem, Camilo Ortiz, Trey Grainger, John A Miller, and William S York. Pgmhd: A scalable probabilistic graphical model for massive hierarchical data problems. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 55–60. IEEE, 2014.

[129] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov chain monte carlo*, 2(11):2, 2011.

[130] Patrick R Conrad, Youssef M Marzouk, Natesh S Pillai, and Aaron Smith. Accelerating asymptotically exact mcmc for computationally intensive models via local approximations. *Journal of the American Statistical Association*, 111(516):1591–1607, 2016.

[131] Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Speeding up mcmc by efficient data subsampling. *Journal of the American Statistical Association*, 2018.

[132] Christian P Robert, Víctor Elvira, Nick Tawn, and Changye Wu. Accelerating mcmc algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5):e1435, 2018.

[133] Michael Braun and Paul Damien. Scalable rejection sampling for Bayesian hierarchical models. *Marketing Science*, 35(3):427–444, 2016.

[134] Brandon Berman. *Asymptotic posterior approximation and efficient MCMC sampling for Generalized Linear Mixed Models*. PhD thesis, UC Irvine, 2019.

[135] Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *2009 IEEE 12th international conference on computer vision*, pages 545–551. IEEE, 2009.

[136] Todd L Graves. Automatic step size selection in random walk metropolis algorithms. *arXiv preprint arXiv:1103.5986*, 2011.

[137] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*, pages 2445–2453, 2013.

[138] Robert E Kass, Bradley P Carlin, Andrew Gelman, and Radford M Neal. Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.

[139] Wan-Kai Pang, Jonathan J Forster, and Marvin D Troutt. Estimation of wind speed distribution using markov chain monte carlo techniques. *Journal of Applied Meteorology*, 40(8):1476–1484, 2001.

[140] Kevin P Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *Technical report (2007)*. URL https://www.cs.ubc.ca/~murphyk/mypapers.html.

[141] Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

[142] Walter R Gilks and Pascal Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

[143] John Geweke and Hisashi Tanizaki. Bayesian estimation of state-space models using the Metropolis–Hastings algorithm within Gibbs sampling. *Computational statistics & data analysis*, 37(2):151–170, 2001.

[144] Leigh J Halliwell. The log-gamma distribution and non-normal error. *Variance: Advancing the Science of Risk*, 2018.

[145] Necdet Batir. On some properties of digamma and polygamma functions. *Journal of Mathematical Analysis and Applications*, 328(1):452–465, 2007.

[146] J Mačys. On the euler-mascheroni constant. *Mathematical Notes*, 94, 2013.

[147] Ross L Prentice. A log-gamma model and its maximum likelihood estimation. *Biometrika*, 61(3):539–544, 1974.

[148] Lauren M Kucirka, Stephen A Lauer, Oliver Laeyendecker, Denali Boon, and Justin Lessler. Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based sars-cov-2 tests by time since exposure. *Annals of Internal Medicine*, 2020.

[149] Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

[150] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, 2014.

[151] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[152] Andrew Gelman, Daniel Lee, and Jiqiang Guo. Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543, 2015.

[153] Team, R Core and others. R: A language and environment for statistical computing, 2013.

[154] GS Rodrigues, David J Nott, and Scott A Sisson. Likelihood-free approximate gibbs sampling. *Statistics and Computing*, pages 1–17, 2020.

[155] Peter Pederson, Danile Dudenhoeffer, Steven Hartley, and May Permann. Critical infrastructure interdependency modeling: A survey of US and international research. *Idaho National Laboratory*, 25:27, 2006.

[156] Steven M Rinaldi. Modeling and simulating critical infrastructures and their interdependencies. In *System sciences, 2004. Proceedings of the 37th annual Hawaii international conference on*, pages 8–pp. IEEE, 2004.

[157] Steven M Rinaldi, James P Peerenboom, and Terrence K Kelly. Identifying, understanding, and analyzing critical infrastructure interdependencies. *IEEE Control Systems*, 21(6):11–25, 2001.

[158] Rae Zimmerman. Social implications of infrastructure network interactions. *Journal of Urban Technology*, 8(3):97–119, 2001.

[159] Donald D Dudenhoeffer, May R Permann, and Milos Manic. Cims: A framework for infrastructure interdependency modeling and analysis. In *Proceedings of the 38th conference on Winter simulation*, pages 478–485. Winter Simulation Conference, 2006.

[160] Earl E Lee II, John E Mitchell, and William A Wallace. Restoration of services in interdependent infrastructure systems: A network flows approach. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6): 1303–1317, 2007.

[161] Nils Goldbeck, Panagiotis Angeloudis, and Washington Y Ochieng. Resilience assessment for interdependent urban infrastructure systems using dynamic network flow models. *Reliability Engineering & System Safety*, 2019.

[162] Leonardo Dueñas-Osorio, James I Craig, Barry J Goodno, and Ann Bostrom. Interdependent response of networked systems. *Journal of Infrastructure Systems*, 13(3): 185–194, 2007.

[163] Isaac Hernandez-Fajardo and Leonardo Dueñas-Osorio. Probabilistic study of cascading failures in complex interdependent lifeline systems. *Reliability Engineering & System Safety*, 111:260–272, 2013.

[164] Min Ouyang and Zhenghua Wang. Resilience assessment of interdependent infrastructure systems: With a focus on joint restoration modeling and analysis. *Reliability Engineering & System Safety*, 141:74–82, 2015.

[165] Wei Koong Chai, Vaios Kyritsis, Konstantinos V Katsaros, and George Pavlou. Resilience of interdependent communication and power distribution networks against cascading failures. In *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, pages 37–45. IEEE, 2016.

[166] Roberto Guidotti, Hana Chmielewski, Vipin Unnikrishnan, Paolo Gardoni, Therese McAllister, and John van de Lindt. Modeling the resilience of critical infrastructure: The role of network dependencies. *Sustainable and resilient infrastructure*, 1(3-4): 153–168, 2016.

[167] X Liu, E Ferrario, and Enrico Zio. Resilience analysis framework for interconnected critical infrastructures. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 3(2):021001, 2017.

[168] Neetesh Sharma, Armin Tabandeh, and Paolo Gardoni. Recovery optimization of interdependent infrastructure: A multi-scale approach. *13th International Conference on Applications of Statistics and Probability in Civil Engineering*, 2019.

[169] MA Di Muro, CE La Rocca, HE Stanley, S Havlin, and LA Braunstein. Recovery of interdependent networks. *Scientific reports*, 6:22834, 2016.

[170] Min Ouyang, Liu Hong, Zi-Jun Mao, Ming-Hui Yu, and Fei Qi. A methodological approach to analyze vulnerability of interdependent infrastructures. *Simulation Modelling Practice and Theory*, 17(5):817–828, 2009.

[171] Yanlu Zhang, Naiding Yang, and Upmanu Lall. Modeling and simulation of the vulnerability of interdependent power-water infrastructure networks to cascading failures. *Journal of Systems Science and Systems Engineering*, 25(1):102–118, 2016.

[172] Myles D Garvey, Steven Carnovale, and Sengun Yeniyurt. An analytical framework for supply network risk propagation: A Bayesian network approach. *European Journal of Operational Research*, 243(2):618–627, 2015.

[173] Nita Yodo and Pingfeng Wang. Resilience modeling and quantification for engineered systems using Bayesian networks. *Journal of Mechanical Design*, 138(3): 031404, 2016.

[174] Seyedmohsen Hosseini and Kash Barker. Modeling infrastructure resilience using Bayesian networks: A case study of inland waterway ports. *Computers & Industrial Engineering*, 93:252–266, 2016.

[175] Chen Chen, Hanghang Tong, Lei Xie, Lei Ying, and Qing He. Fascinate: Fast cross-layer dependency inference on multi-layered networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 765–774. ACM, 2016.

[176] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009.

[177] Mohammad Al Hasan and Mohammed J Zaki. A survey of link prediction in social networks. In *Social network data analytics*, pages 243–275. Springer, 2011.

[178] Susan L Cutter, Bryan J Boruff, and W Lynn Shirley. Social vulnerability to environmental hazards. *Social science quarterly*, 84(2):242–261, 2003.

[179] Wesley E Highfield, Walter Gillis Peacock, and Shannon Van Zandt. Mitigation planning: Why hazard exposure, structural vulnerability, and social vulnerability matter. *Journal of Planning Education and Research*, 34(3):287–300, 2014.

[180] Walter Gillis Peacock, Shannon Van Zandt, Yang Zhang, and Wesley E Highfield. Inequities in long-term housing recovery after disasters. *Journal of the American Planning Association*, 80(4):356–371, 2014.

[181] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[182] Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160, 1972.

[183] M Shinozuka, A Rose, and RT Eguchi. Engineering and socioeconomic impact of earthquakes: An analysis of electricity lifeline disruptions in the New Madrid Area Monograph 2. *Multidisciplinary Center for Earthquake Engineering Research, Red Jacket Quadrangle, State University of New York at Buffalo, Buffalo, NY*, 14261, 1998.

[184] Andrés D González, Airlie Chapman, Leonardo Dueñas-Osorio, Mehran Mesbahi, and Raissa M D'Souza. Efficient infrastructure restoration strategies using the recovery operator. *Computer-Aided Civil and Infrastructure Engineering*, 32(12):991–1006, 2017.

[185] Census Bureau. *Census Data*. Department of Commerce, Washington, D.C., 2018. URL https://www.census.gov/data.html.

[186] CDC. *Social Vulnerability Index (SVI)*. Department of Health and Human Services, Washington, D.C., 2016. URL https://svi.cdc.gov/data-and-tools-download.html.

[187] FEMA. *Hazus-MH 2.1 User Manual: Earthquake Model Technical Manual, Washington, DC.*, 2013.

[188] Min Ouyang, Leonardo Dueñas-Osorio, and Xing Min. A three-stage resilience analysis framework for urban infrastructure systems. *Structural safety*, 36:23–31, 2012.

[189] Yi-Ping Fang and Enrico Zio. An adaptive robust framework for the optimization of the resilience of interdependent infrastructures under natural hazards. *European Journal of Operational Research*, 276(3):1119–1136, 2019.

[190] Tomas Lagos, Rodrigo Moreno, Alejandro Navarro, Mathaios Panteli, Rafael Sacaan, Fernando Ordonez, Hugh Rudnick, and Pierluigi Mancarella. Identifying optimal portfolios of resilient network investments against natural hazards, with applications to earthquakes. *IEEE Transactions on Power Systems*, 2019.

[191] A PROCLAMATION. Presidential proclamation–critical infrastructure security and resilience month, 2019. 2019.

[192] Lawrence V Snyder, Maria P Scaparra, Mark S Daskin, and Richard L Church. Planning for disruptions in supply chain networks. In *Models, methods, and applications for innovative decision making*, pages 234–257. INFORMS, 2006.

[193] Ricardo Rubio-Barros, Diego Ojeda-Esteybar, Osvaldo Añó, and Alberto Vargas. *Combined operational planning of natural gas and electric power systems: State of the art*. InTech, 2010.

[194] M Shahidehpour and Z Li. White paper: Long-term electric and natural gas infrastructure requirements. *llinois Institute of Technology, Tech. Rep*, 2014.

[195] PN Biskas, NG Kanelakis, A Papamatthaiou, and I Alexandridis. Coupled optimization of electricity and natural gas systems using augmented lagrangian and an alternating minimization method. *International Journal of Electrical Power & Energy Systems*, 80:202–218, 2016.

[196] Jin-Zhu Yu and Hiba Baroud. Modeling uncertain and dynamic interdependencies of infrastructure systems using stochastic block models. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 6(2), 2020.

[197] Ece Sanci and Mark S Daskin. Integrating location and network restoration decisions in relief networks under uncertainty. *European Journal of Operational Research*, 279(2):335–350, 2019.

[198] Bismark Singh, David P Morton, and Surya Santoso. An adaptive model with joint chance constraints for a hybrid wind-conventional generator system. *Computational Management Science*, 15(3-4):563–582, 2018.

[199] François V Louveaux and Rüdiger Schultz. Stochastic integer programming. *Handbooks in operations research and management science*, 10:213–266, 2003.

[200] Suvrajeet Sen. Algorithms for stochastic mixed-integer programming models. *Handbooks in operations research and management science*, 12:515–558, 2005.

[201] Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

[202] Daniel Delahaye, Supatcha Chaimatanan, and Marcel Mongeau. Simulated annealing: From basics to applications. In *Handbook of Metaheuristics*, pages 1–35. Springer, 2019.

[203] Darrall Henderson, Sheldon H Jacobson, and Alan W Johnson. The theory and practice of simulated annealing. In *Handbook of metaheuristics*, pages 287–319. Springer, 2003.

[204] Lihua Sun, L Jeff Hong, and Zhaolin Hu. Balancing exploitation and exploration in discrete optimization via simulation through a gaussian process-based search. *Operations Research*, 62(6):1416–1438, 2014.

[205] Iain Dunning, Joey Huchette, and Miles Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.

[206] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[207] Gurobi Optimization. Gurobi optimizer 9.0. *Gurobi: http://www. gurobi. com*, 2019.

[208] Qiling Zou and Suren Chen. Enhancing resilience of interdependent traffic-electric power system. *Reliability Engineering & System Safety*, 191:106557, 2019.

[209] Vipin U Unnikrishnan and John W van de Lindt. Probabilistic framework for performance assessment of electrical power networks to tornadoes. *Sustainable and Resilient Infrastructure*, 1(3-4):137–152, 2016.

[210] U.s. energy information administration. https://www.eia.gov/electricity/annual/html/epa_08_04.html, 2018. Accessed: 2020-09-17.

[211] Center point energy. https://www.centerpointenergy.com/en-us/Services/Pages/natural-gas-electricity-cost-comparison.aspx?sa=mn&au=bus, 2018. Accessed: 2020-09-18.

[212] Min Ouyang and Leonardo Duenas-Osorio. Multi-dimensional hurricane resilience assessment of electric power systems. *Structural Safety*, 48:15–24, 2014.

[213] Chemical engineering portal. https://missrifka.com/equipments/compressor/centrifugal-compressor-power-calculation.html, 2020. Accessed: 2020-09-18.

[214] U.s. energy information administration. https://www.eia.gov/tools/faqs/faq.php?id=107&t=3:%20efficienty%200.4, 2020. Accessed: 2020-09-17.

[215] Mathaios Panteli, Cassandra Pickering, Sean Wilkinson, Richard Dawson, and Pierluigi Mancarella. Power system resilience to extreme weather: fragility modeling, probabilistic impact assessment, and adaptation measures. *IEEE Transactions on Power Systems*, 32(5):3747–3757, 2016.

[216] Harvey Thurm Taylor, Bill Ward, Mark Willis, and Walt Zaleski. The saffir-simpson hurricane wind scale. *Atmospheric Administration: Washington, DC, USA*, 2010.

[217] Cliff Grigg, Peter Wong, Paul Albrecht, Ron Allan, Murty Bhavaraju, Roy Billinton, Quan Chen, Clement Fong, Suheil Haddad, Sastry Kuruganty, et al. The ieee reliability test system-1996. a report prepared by the reliability test system task force of the application of probability methods subcommittee. *IEEE Transactions on power systems*, 14(3):1010–1020, 1999.

[218] Flow control network. https://www.flowcontrolnetwork.com/instrumentation/flow-measurement/article/15554850/defining-maximum-gas-pipe-velocity, 2011. Accessed: 2020-09-17.

[219] Howard Raiffa and Robert Schlaifer. *Applied statistical decision theory*. Harvard University, Boston, MA, 1961.

[220] Lina Sela and Saurabh Amin. Robust sensor placement for pipeline monitoring: Mixed integer and greedy optimization. *Advanced Engineering Informatics*, 36:55–63, 2018.

[221] Adam N Elmachtoub and Paul Grigas. Smart” predict, then optimize”. *arXiv preprint arXiv:1710.08005*, 2017.