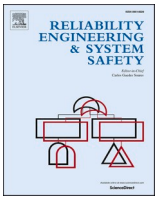
Reliability Engineering and System Safety 208 (2021) 107348

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



Power outage prediction for natural hazards using synthetic power distribution systems

Chengwei Zhai *, Thomas Ying-jeh Chen, Anna Grace White, Seth David Guikema

Industrial & Operations Engineering Department, University of Michigan, Ann Arbor, MI 48109, USA

ARTICLE INFO

Keywords:
Distribution system
Power system reliability
Data models
Power outage estimation

ABSTRACT

Power outage prediction for natural hazards usually relies on one of two approaches, statistical models or fragility-based methods. Statistical models have provided strong predictive accuracy, but only in an area-aggregated manner. Fragility-based approaches have not offered strong prediction accuracy and have been limited to systems for which system topology or performance models are available. In this paper, we create an algorithm that (1) generates a synthetic power system layout for any U.S. city based only on public data and then (2) simulates power outages at the level of individual buildings under hazard loading using fragility functions. This approach provides much more localized, building-level estimates of the likelihood of losing power due to a natural hazard. We validate our model by comparing the network properties and power outage events based on our approach with data from a real power system in Ohio. We find that our model relies on less input data comparing to statistical learning approaches yet can make accurate predictions, provided accurate fragility curves are available.

1. Introduction

Power outages regularly cause large economic losses, impact other critical infrastructure systems and significantly disrupt daily life. The leading cause of large-scale power outages in the U.S. are severe weather events, though other natural hazards such as earthquakes can lead to significant outage events as well [1]. By some estimates, 75% of power outages are either directly caused by weather-induced faults, or indirectly through failure of aging equipment exposed to significant weather events [2]. For example, in 2012, a powerful derecho struck the Midwestern United States and caused 4.2 million customers across 11 states to lose power. Power restoration took up to 10 days for some areas [3].

Power outage estimation can help power utilities, the managers of other critical infrastructure that are dependent on power, governments, and private organizations with both short-term event response and longer-term resilience planning. In the short term, pre-event power outage estimation can help utilities better plan their response and thus better balance cost and restoration speed. For large-scale outage events, utilities rely heavily on outside personnel and restoration material through mutual aid agreements. These resources are critical in restoring power quickly but are very costly, with restoration costs in the tens of millions of dollars per day for larger utilities [4]. If a utility underestimates the outages from a forecast weather event and does not bring in enough mutual aid, their customers face prolonged outages. On the other hand, if a utility overestimates outages and brings in more mutual aid than is needed, they incur unnecessarily high costs which must then be borne by either rate-payers or corporate owners or shareholders. In the longer term, being able to estimate the likelihood of power outages at the individual building level from both weather events and other hazards such as earthquakes and floods can help utilities better understand where system strengthening may be needed. Building-level estimates of the likelihood of losing power can also help businesses and homeowners better understand and improve their resilience locally through considerations of back-up power, power disruption insurance (for a commercial entity), and other measures.

There are two main approaches for estimating power outages in the literature, statistical models and fragility-based models [5,6]. These will be discussed in more details in Section 2, but briefly, statistical models have been implemented successfully in practice and can provide accurate estimates of power outages due to a forecast weather event at a spatial unit scale (e.g., census tract, zip code, or county). However, to date, they have not provided building-level estimates. On the other hand, fragility-based approaches do provide building-level outage estimates but they (1) rely on data about the power system layout that is not

* Corresponding author.
E-mail addresses: cwzhai@umich.edu (C. Zhai), tyjchen@umich.edu (T.Y.-j. Chen), agracew@umich.edu (A.G. White), sguikema@umich.edu (S.D. Guikema).

https://doi.org/10.1016/j.ress.2020.107348
Received 11 May 2020; Received in revised form 2 October 2020; Accepted 4 October 2020
Available online 24 November 2020

usually available, (2) have not shown strong accuracy to date and (3) can be computationally difficult and require a non-negligible amount of data collection effort to scale up to national or global level analysis.

The work in this paper significantly advances the second of these approaches – fragility-based outage estimation – by developing a new framework to accurately estimate the probability of losing power at the individual building level using only publicly available data. The fragility-based approach simulates how each component of the system fails given fragility curves and determine the functionality of the system given these failure components. However, there is a fundamental challenge. The topology of power distribution systems is not publicly available in the U.S. due to security concerns. Because of this, we first develop a method to generate a synthetic power distribution system based on only publicly available data. The method is generalizable throughout the US and can be scaled from the city level to a larger spatial level, such as county level or state level. We then develop a method to simulate power outages at the building level based on this synthetic system in combination with hazard loading and fragility information. This approach can be used to simulate many types of natural disasters, e.g. hurricanes, flooding, and earthquakes, provided that relevant hazard loading maps and fragility curves are available.

The structure of this paper is as follows. We first discuss previous research on power outage prediction, synthetic power grid generation, and infrastructure vulnerability analysis. We then introduce our methodology to generate a synthetic distribution system and how to simulate its performance under hazard loading. Next we validate our synthetic network layouts against actual distribution systems and at last we test the outage simulations against historical large-scale power outage events to demonstrate the accuracy and functionality of the system.

2. Literature review

Before giving an overview of the related literature, one point of clarification is needed. Most outages due to severe weather events occur in the power distribution system, while for earthquakes there is often also damage to substations (and sometimes the transmission system and generators) [7]. In comparison to the power transmission system, the power distribution system, which delivers power from local substations to each customer, is more vulnerable and more likely to be damaged and cause customers to lose power during adverse weather events [8]. For the distribution system, when a line breaks, a pole falls, or a distribution substation fails, customers that are downstream of the failed devices will be isolated and lose power unless there is an alternate set of lines that can serve that customer [9]. This redundancy is rare in power distribution systems; most power distribution systems are dominantly radial in design [10,11]. The method we present in this paper focuses on low voltage substations and the distribution system, though it can incorporate the transmission system as well provided that network layout information is available. Our literature review covers both systems. We first review previous work on statistical approaches for power outage prediction. Then we review methods of generating synthetic power systems followed by a review of fragility curves and outage estimation approaches based on fragility curves.

2.1. Statistical power outage predictions

One of the main approaches in both the research literature and in use in practice for estimating power outages are statistical methods. Many, though not all, of these studies focus on tropical cyclones [5,12–18]. Others focus more broadly on a range of weather conditions [19]. These approaches use data from past events together with a wide array of explanatory variables to develop statistical and machine learning models to predict power outages due to a forecast weather event.

The statistical models used for outage and damage forecasting have varied from relatively simple generalized linear and generalized additive models in early work [13,20] to regression trees [15] to ensembles of trees and other machine learning methods more recently [5,16,18,19,21,22]. One consistent challenge in using statistical approaches for outage forecasting is zero-inflation of the outage data, meaning that even for significant adverse events, many more areas experience zero outages than experience outages if the spatial units used are small. Guikema and Quiring developed a two-stage process that combines a classification model and a regression model to first predict in a given area whether power outages will occur or not, for those areas that they forecast will experience outages we then proceed to estimate the severity of the event (e.g. number of outages) [23]. Shashaani et al. (2018) and Kabir et al. (2019) developed three-stage approaches that introduced a new stage to predict the severity class of power outages followed by a quantile regression forest to provide a probabilistic prediction [18,19].

Statistical outage forecasting models have focused on severe weather events, and the input data used has consequently focused on features that may help predict the number of outages. This has included weather forecast information from numerical weather forecasting models [19,24] or, for hurricanes, from hurricane wind field models [5,16,20]. It has also included information about the assets exposed to the hazard such as the numbers of poles, transformers, and wire spans in each spatial unit [17–20]. In addition, it has also included a range of features that describe local geography, plant species, utility vegetation management, pre-storm soil moisture levels, and plant species in each area [17,25]. These features have been found to offer improved predictive accuracy. Other approaches, particularly those of Guikema et al. and Nateghi et al. have sought to use a reduced set of input features to predict power outages due to hurricanes [16,26]. The advantages of easier implementation in practice and potentially greater generalizability, but at the cost of some loss of predictive accuracy. Regardless of the details, all of these approaches provide outage estimates at the level of an aggregated spatial level, which may vary from relatively small (e.g., 5 km by 5 km square) grid cells to census tracts, counties, or utility operating districts. They do not provide building or facility-level estimates.

2.2. Fragility curves and fragility curve based methods

The other main approach for estimating loss of power and power system damage due to natural hazards is one based on fragility curves. This is the approach implemented in HAZUS, a natural hazards loss estimation software package supported by the U.S. Federal Emergency Management Agency (FEMA). A fragility curve is a function giving the probability of a particular infrastructure asset or building reaching or exceeding a specific damage state given a quantified disaster intensity metric. For example, a fragility curve could give the probability that a substation is in each of four damage severity states as a function of the ground shacking, measured by peak ground acceleration, due to an earthquake at that substation's location. One of the more widely used set of fragility curves are those in HAZUS for damage to infrastructure components and buildings from earthquakes, wind events, floods, and tsunamis [27,28].

Fragility functions are of course hazard-dependent, and some hazards are more well-studied than others. Fragility functions for power system components for seismic events have been particularly well-studied. Power system components can be divided into micro-components (e.g., coil support, circuit breaker, transformer) and macro-components (a combination of micro-components). Vanzi developed some of the earliest fragility functions for electric power system components for seismic events based on a functional form given by a cumulative lognormal distribution [29]. HAZUS includes fragility curves for a broader set of power system assets, including generation plants, substations, and distribution. For example, HAZUS classifies substations into low voltage, medium voltage and high voltage. For each voltage level, the substation can be anchored or unanchored. It then describes the damage of electric power substation from earthquake with

five different severity states and provides the lognormal parameters of the probability the substation exceeds each damage state given the peak ground acceleration.

Fragility curves for strong wind events are not as developed as for seismic events, though significant progress has been made in the last several years. The most critical asset for wind events is the poles in the system as these are the major locations of failures and thus cause of outages during wind events. An early approach is that of Han et al. which used a structural reliability model to estimate a fragility curve for power poles, used this to formulate a prior probability distribution, and then updated this with observed pole failure data [30]. However, they were hampered by insufficient observations of pole states after events. Mohammadi et al. provide a more recent and comprehensive study on utility poles fragility curves for strong wind events [31]. They consider pole age, conductor area, height and wind direction as variables in the lognormal fragility function. However, none of this prior work incorporates the impact of trees falling and shedding branches onto power lines or power poles in their functions, and this can be a major cause of failure [16,32].

A set of fragility curves on its own is not a complete model for estimating power outages. The fragility curves must be coupled with a method for simulating realizations of damage states of the set of assets and then for translating each of these realizations into an estimate of which customers lose power. If the actual power distribution system layout is known, then this is a relatively straight-forward task. Typically, a Monte Carlo simulation is used to generate N replications of asset damage states, where one replication includes a damage state for each asset in the system. Then either a power flow or, more often for a distribution system level analysis, a connectivity-based model is used to determine whether or not each customer has power based on the set of assets damaged between that customer and the substation serving them. If there is redundancy in the system, then a power flow analysis is likely necessary instead of a simple connectivity model [33].

A challenge is that the actual power system layout is rarely available at the distribution level, which is the level that is most needed for outage estimation. Power utilities do not share this data publicly due to security concerns. While specific researchers do get access to the layouts for specific utilities for specific projects, this to date has made it very difficult to apply fragility-based approaches widely. One way around this has been to assume a system topology. For example, Winkler et al. (2010) assumed that all roads in a city had a power line and then used this to simulate outages. This, however, does not approximate the topology of actual distribution systems well, particularly in how each building is connected through the network to a substation. An alternative, the one we develop further in this paper, is to generate synthetic layouts that better represent the layout of power systems.

2.3. Synthetic power system generation

There has been some previous work on synthetic power grid generation. Most of this previous research has focused on the transmission system [34–37]. For example, Birchfield et al. propose a method to generate synthetic transmission systems together with validation criteria for these systems [34]. They first place high voltage substations with a clustering algorithm considering the spatial distribution of customers. Then they add in transmission lines that would meet power flow constraints. They test their model with a 2000-bus public test case. Soltan and Zussman (2016) present a Geographical Learner and Generator Algorithm (GLGA) to generate synthetic transmission networks similar to a given network comparing similar structural and spatial properties such as average path length, clustering coefficient, degree distribution of the nodes, and length distribution of the lines. Some recent work also tries to model synthetic distribution networks [38,39]. Pisano et al. use georeferenced information and other publicly available open data to estimate the energy consumption of a region. Based on locations of primary substations and territory segmentation they create the layout of the distribution system which can be used further in optimization studies. However, these generated synthetic distribution networks cannot serve as a good representation of the actual system when determining power outages during natural disasters because they, (1) lack critical details such as the locations of poles undergrounding and (2) lack validation against actual distribution systems.

For the distribution grid, considerable work has been focused on optimal layouts of distribution networks [40-42].Miranda et al. (1994) used a genetic algorithm to plan the placement of new distribution network to expand from an existing system. They assumed the network to be radial and optimized the system to minimize new facility installation costs and operation costs under constraints such as power flow, voltage drop, and power demand etc. Their approach starts with possible sites for substations and potential power line locations and uses the genetic algorithm to find an optimal solution for this binary integer optimization problem. Valenzuela et al. (2019) used a Minimum Spanning Tree model to create a distribution network with georeferenced data (e.g. customers' locations, street point positions). They focus on the optimal allocation of distribution transformers and assumed undergrounding lines only, a situation that would be quite uncommon in the U.S. The goal is to create a distribution network that minimizes the total load shed during extreme events. Overall, these approaches provide starting points, but there does not yet exist an approach that allows us to create a synthetic power distribution network that is representative of power distribution systems and then to simulate hazard-induced failures and estimate the probability of loss of power at the individual building level. This is the challenge we address in this paper.

3. Methodology

3.1. Assumptions and information needed

Even though distribution network layouts are not publicly available, two critical aspects that help define a distribution network can be acquired or can be proxied. One of these is the location of customers (power meters). We know that every building in a developed nation can receive electric power so we can assume the locations of meters are the locations of buildings and that each building must thus be served by the synthetic system. This is not an exact representation of the number of customers because, for example, some multi-unit buildings have multiple meters. However, the spatial distribution of building locations is a good proxy for the actual customer locations for the purpose of creating locations for synthetic power lines. Such information can be retrieved from building footprints that are publicly available for most cities in the US.

Another critical component in developing a synthetic power distribution system is the location of distribution substations. The locations of distribution substations are publicly available from open-source map platforms.[1] We can view these substations as power supply points that deliver power to the distribution feeders leading to each customer.

Three key assumptions underlie our approach for generating synthetic power distribution systems. These assumptions are:

1) All powerlines are within roads' rights-of-way,
2) the system is designed in manner that at least approximately the least-cost method of connecting all customers in terms of total line miles, and
3) the network is radial. That is, it has a tree-structure.

None of these are strictly always true for real systems. However, they are true for large portions of many U.S. power distribution systems [43], and, as we show below, they allow us to achieve our goal of accurately

[1] e.g. *overpass-turbo.eu*.

estimating the likelihood of power outages at the individual building level based on only publicly available information. We will elaborate further on these below. With these two sets of locations, customers and substations, and three key assumptions, we can create a synthetic distribution power network to simulate power outages under extreme hazard events.

We focus our approach on the major power infrastructure components that could be damaged by natural hazards. That is, rather than trying to model every component, we focus on the asset classes that are the main sources of loss of power and that correspond to the available fragility curves, namely substations, poles, and power lines. Different types of power system components have different responses to natural disasters. Substations, which transform voltage from high to low, are more likely to be damaged during earthquakes and flooding events than during wind events. In a radial system, damage to substations may lead to loss of entire feeders (a set of distribution lines serving one portion of a substation's service area), creating an outage that impacts many customers simultaneously. Poles are more likely to be damaged during strong wind events and, if there is liquefaction, during earthquakes. Pole damage in a radial system cuts power to all downstream customers. Above-ground power lines can be damaged from falling trees and limbs and blown debris, and below-ground lines can be damaged due to flooding. Fragility curves, e.g., those in HAZUS, often aggregate the line damage probabilities and assign it to the closest upstream pole. We adopt this approach and do not assign separate failures probabilities to lines.

We create network layouts for a power distribution system, each of which describes how each customer gets power from distribution substations through power lines. Each layout is a graph with the customers, substations, and poles as vertices and the power lines as the edges. This introduces three questions we need to answer: 1) which substation each customer gets power from, 2) the route from each substation to each customer, and 3) whether the power lines in each location are overhead or underground.

As discussed above, one of our key assumptions is that power lines are strictly along the roads. Based on our observation from the actual distribution layout in our case study system, that of Franklin County, Ohio, the average distance from a powerline to the nearest road is less than 100 m for 92% of power lines are within 100 m to roads. We calculate the Euclidean distance from each end point of each segment of power line polyline to the closest road intersections and average the value for each polyline to make this calculation.

One of the major advantages of our approach is the low requirement of data collection in comparison to statistical learning approaches [18]. We use only open-source data, i.e., road layouts, building locations, partial building information, and substation locations, to generate the power system network. Road layouts are shapefiles available from United States Census Bureau at county level.[2] Power demands locations are the coordinates of each customer within the city/region boundaries. We extract these coordinates from any given city's building footprints and approximate each customer location with the centroid of each building.[3] For supply points, we focus on the distribution system because this is the portion of the system most vulnerable to get damaged from many types of hazards [8]. We use open-source map query website *overpass.turbo* to identify all the substation locations and download them and we use *Zillow.com* to retrieve building information. With this set of information, we can then create our synthetic power network that provides power connectivity from substations to all customers.

[2] E.g., https://catalog.data.gov/dataset/tiger-line-shapefile-2015-county-franklin-county-oh-all-roads-county-based-shapefile.

[3] E.g. https://geo.btaa.org/catalog/6a0036b36c004d53b747d322265df751_1.

3.2. Network generation methodology

In this section, we introduce the process of generating a synthetic power system layout with different approaches and validate those layouts. In general, the process can be divided into 6 steps as the flowchart in Fig. 1 shows.

3.2.1. Customer clustering

The first step in generating the distribution system is to create a cluster of customers served by each substation. This process answers our first question in synthetic distribution network generation – which substation a customer gets power from. In the U.S., where distribution systems are typically radial, each customer is served by only one substation. In practice, many assign customers to substations based on closest euclidean distance. However, this is not necessarily the case. Sometimes for a customer, the closest substation following the path of power lines is not the substatuion with the closest euclidian distance. For example, topological effects, water bodies, and other considerations can lead a customer to be served by a substation other than the closest Euclidean distance substation. The assumption that customers are served by the closest substation by euclidian distance can lead to substantially different results for risk assessment purposes. It eliminates the possibility the customer is actually served by another substation which experiences a very different damage condition. It is crucial to determine the potential substation service territory based on network distance, not euclidian distance.

Our method uses the following approach to determine the service territory of each substation. We first calculate the distances from each customer to all the substations along the road network. Then for each customer, we find the closest substation based on network distances. This gives us a basic customer cluster for each substation. This is helpful to determine the service substation for customers that are only close to one substation. However for customers that have a similar network distances to multiple substations, we generate multiple realizations of the layout that allow these customers to be in different clusters. This acknowledges that we are uncertain about which substation they are served by. We define c_i as the cluster that customer i belongs to and $d_n(i,j)$ as the network distance between customer i, $i \in [1,N]$ with substation $j \in [1,M]$. Then we have

$$c_i = j, \ if \ \frac{d_n(i,j) - \min\limits_{j=1:\mathrm{M}} d_n(\mathrm{i},\mathrm{j})}{\min\limits_{j=1:\mathrm{M}} d_n(\mathrm{i},\mathrm{j}))} < b \tag{1}$$

where b is the threshold (10–20%) to constrain the relative distances from customers to other substations comparing to their closest substation. This means that customers on the boarders of clusters can appear in different substation clusters in different realizations of the network layout. As an example shown in Fig. 2, where there are two substations A and B. Orange customers and yellow customers are served by substation A and B respectively as they are very close to their substation. The two green customers have a similar distances to both substations, so they are considered to be in both substation clusters. This does not mean green customers are served by two substations at the same time but they are simulated independently in each substation cluster that has them,

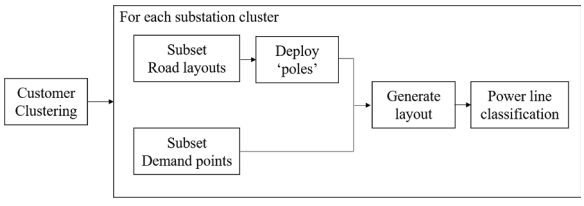


Fig. 1. The six steps of our synthetic power distribution system network generation algorithm.

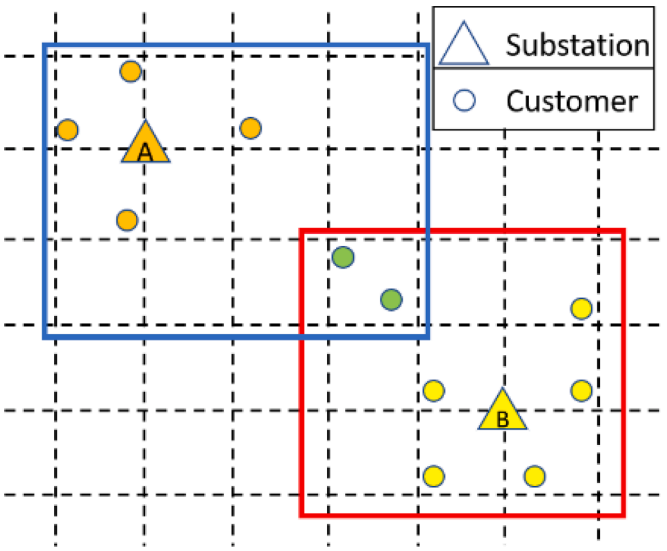


Fig. 2. Illustration of border customers and their assignment to different substatiuons.

generating multiple network layouts that form an ensemble. As the output from the simulation, depending on the purpose of the study, we can choose to report the average probability across the ensemble or the highest probability across the ensemble of the green customers losing power from substation cluster A and B.

3.2.2. Create reduced problem

Instead of creating a fully connected distribution network that connects all the substations and customers, we can create multiple reduced distribution networks given each substation cluster. This simplifies computation and better reflects the layout of real substations. For each substation cluster, we generate a buffer that is slightly larger than the spatial locations of customers. As an example, in Fig. 2, the two substation clusters A and B, along with their customers and the roads within each square buffer compose two independent reduced problems. Then for each reduced problem, we create connectivity for each customer to get power from the substation. The benefit from creating reduced problem is that it can largely reduce the computational effort relative to running algorithms on the entire system and it is a reasonable relaxation to the original problem as the actual system is typically radial and separated into substation feeders in the US.

3.2.3. Generate poles and road segmentation

Each customer in a typical power distribution system is connected to the substation through a combination of overhead lines (required utility poles) and underground lines. Therefore, we need to split the roads into segments such that the distance between two nodes is similar to the actual distance between poles. For each road's polyline, we define the coordinates of all the nodes of the polyline as the sequence as $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$. We define these nodes as power nodes. The point (x_1,y_1) is the beginning node of the road segment, and (x_n,y_n) is the ending node. We define the Euclidean distance between point i and j as $d_e(i,j)$. The sequence of the polyline can be reversed. We assume the distance between poles to be a constant d_{pole} , and we assume this is 40 m based on our observations of the pole distribution of actual systems. However, d_{pole}, can be changed for specific systems if information is available that suggests a different span length for overhead lines. We define S as the set containing all the power nodes we created from the road shapefile.

The algorithm is as follows. We start by putting (x_1,y_1) into S. If the distance between (x_1,y_1) and (x_2,y_2) is less or equal to d_{pole}, we put (x_2,y_2) into S. If the distance is larger than d_{pole}, we split the line between (x_1,y_1) and (x_2,y_2) evenly and the number of nodes added is determined by the floor of $d_e(1,2)/d_{pole}$. For example, if $d_e(1,2) = 50m$, we will add in one point $(x_{1.1},y_{1.1})$ in the middle of the polyline. We then put in nodes $(x_{1.1},y_{1.1})$ and (x_2,y_2) into S. As a result, S should contains all the original nodes from road polylines and new nodes created when the distance between neighboring nodes are further than d_{pole}. If a power node belongs to an overhead power line, it will be viewed as a utility pole which can be exposed and damaged by wind or earthquake-induced liquification. A power feeder node is defined as the power node that is used to connect to at least one customer.

3.2.4. Generate distribution network

The next step is to create connectivity within each cluster to deliver power from substations to customers. We compare three different models to accomplish this goal: Steiner tree, K-mean clustering Steiner tree, and shortest path.

The first method (Algorithm 1) we consider is to connect all customers and the substation in a minimum cost manner. In another words, we want to find the tree with the minimum cost that connects all the nodes of interests on an undirected graph. This is the problem of finding the Steiner tree on the graph [44]. Our base network is the road layouts as we constrain our power lines to be along roads. Our important vertices are power feeders, where one or more buildings receive power. We calculate the closest power node for each building, and those power nodes with at least one customer near it will be considered as important nodes. The Steiner tree problem is a well-studied NP-hard problem, and many approximation algorithms have been created to reduce the difference from the optimal Steiner tree to the approximation solution [45]. We use an approximation algorithm created by Takahashi and Matsuyama [44]. We define the original road network as an undirected graph $G = (N,E)$, consisting the set N of all the potential power nodes we created from 3.2.3 and the set E of all the road segments that connect the power nodes. We define another undirected graph $G' = (N', E')$ to be the Steiner tree we are looking for, $N' \in N$, $E' \in E$. The weight of each edge is the length of the road segment. We define set $N_I \in N$ of all the power feeders. The algorithm is as following.

The second method (Algorithm 2) we examine is called K-mean clustering Steiner tree. The intuition of this method is trying to imitate the development progress of communities. We assume that the development of each substation cluster's distribution system begins with building major power lines from the substation to each neighborhood (i.e., grouping of buildings). Then the power lines within each group of buildings are added, connecting to the original main power line. The algorithm changes as following.

The third method we examine to connect each building to the substation is a shortest weighted path approach. If we have power feeders $(n_1,n_2,n_3\ldots)$ and the substation power node n_s, then we find the shortest path from each power feeder node to a substation power node and include the path into G'. In this way, all the buildings are connected to the substation most efficiently. This method can be viewed as a special case to the second method when the number of clusters is equal to the number of customers.

Each of these three methods generates a distribution layout for each substation cluster. The distribution network connects each customer to a power node, which is connected to the substation. This connectivity

Algorithm 1
Steiner tree to generate distribution layout.

Step 1: Select a random vertex $s \in N_I$, and find a vertex $t \in N'$, $s \neq t$ that gives the shortest weighted path e_{st} to s. Add e to E' and all the vertices in e_{st} to N'. This gives a starting tree that connects s and t. Remove s and t from N_I.

Step 2: Search from all the vertices in N_I and find a vertex u such that the weighted path e_u from u to G' is the shortest. We then add e_u and all the vertices in e_u to N'. We then remove u from N_I.

Step 3: Repeat Step 2 until all the vertices in N_I have been connected to G' and N_I becomes empty.

Algorithm 2

Steiner tree with K-mean clustering to generate distribution layout.

Step 1: Spatially cluster customers within each substation cluster with a K-mean algorithm and determine the best number of clusters based on silhouette score. We use the closest power node to each cluster center in G as our centers for communities, i.e. n_1, n_2, n_3.

Step 2: Then connect the substation's closest power node n_s to each substation cluster's power nodes n_1, n_2, n_3 with the shortest weighted path on G. Then add these vertices and arcs in G'. Remove substation power nodes and community cluster power nodes from N_I.

Step 3: Apply the Steiner tree algorithm to include all the important nodes and paths into G'.

pathway is critical in the simulation step of our algorithm because we assume that if a customer is connected to a substation, they can receive power. That is, we ignore power flow constraints in our radial power distribution system. Therefore, the distance from the substation to a customer is crucial in determining the customer's probability of losing power. The further the customer is to the substation, the more components (e.g., poles) there likely are between the customer and the substation, increasing the possibility of a disruption on the distribution path to cut off power to the customer. Therefore, one of the validation criteria we use below is a comparison of network distance from each customer to the substation in our synthetic network with that in the actual distribution network.

Simple examples of the three methods are shown in Fig. 3. Green dots are power feeders that deliver power from the distribution network to customers, and the triangle is the substation. The goal is to connect all of them on the network grid with certain rules. We assume each dashed line represent roads, and we implement the three methods we proposed to solve the same problem. The three plots in the middle give examples of intermediate steps while solving the problem with each method. The orange lines are expected to be the next power lines to be added into the prior solution represented by blue lines. The three plots on the right are the final solution from each network generation method. The two red stars in the "ST+Kmean" plot are cluster centers.

3.2.5. *Overhead/underground power lines classification*

It is critical to determine whether each power line is overhead or underground because overhead and underground lines have substantially different vulnerabilities to hazards. We use a random forest classifier to estimate whether each line is overhead or underground, and the

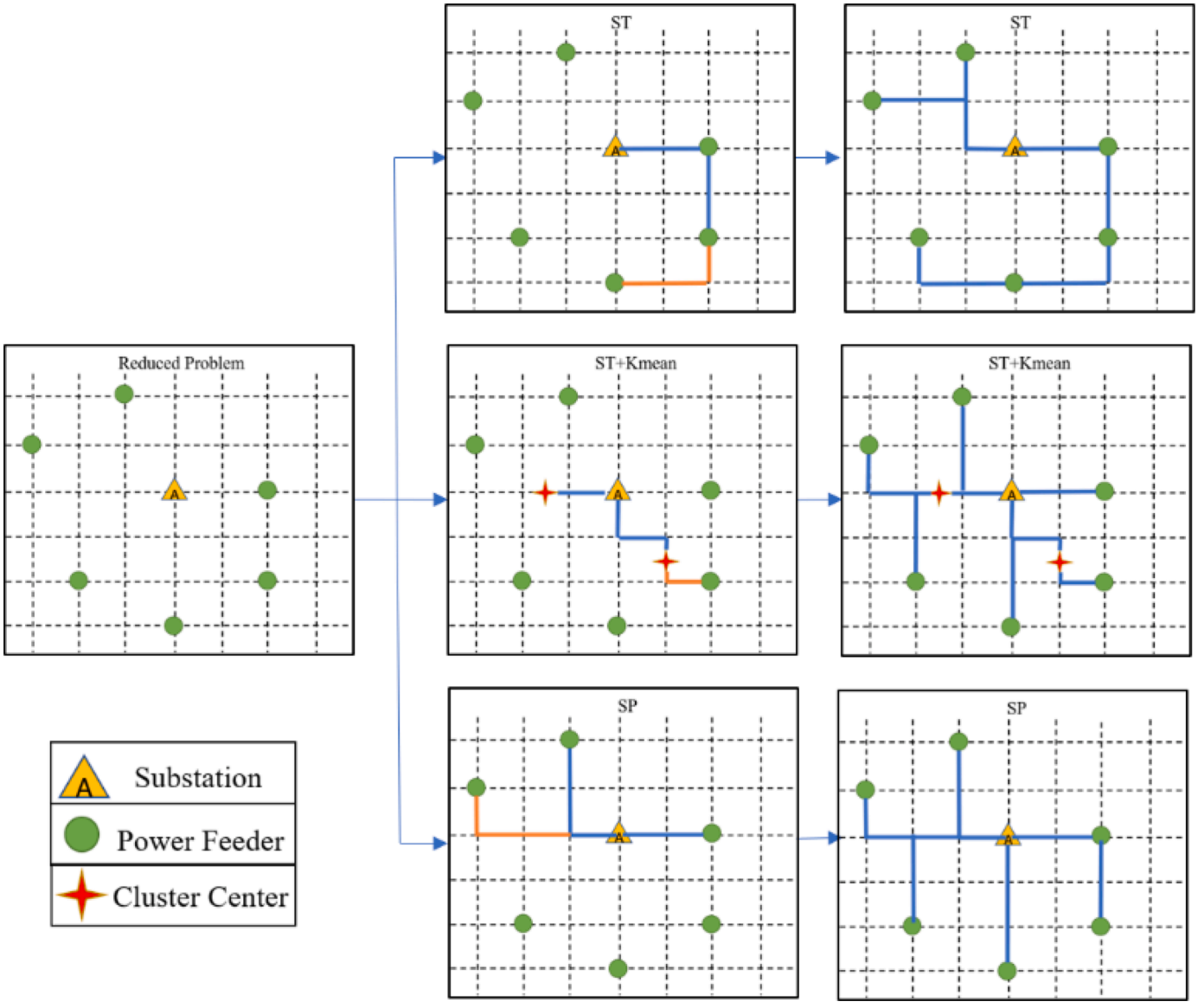


Fig. 3. Illustrative examples of the three network connectivity generation algorithms.

flowchart of this classification process for each substation cluster is shown in Fig. 4. This approach uses the actual status – overhead vs. underground – of each line segment in the actual system we have data for together with housing characteristics of the area around the lines to develop a predictive machine learning classifier. The assumption here, based on observations working with data from multiple utilities, is that underground lines are more likely in certain types of areas such as those with newer homes with higher values.

We start by collecting real estate information for the area from *Zillow.com*. Other sources of real estate information could be used, but this website offers a convenient, free source of data. We take a few factors into consideration as predictive variables and train a statistical learning model to classify the line type that connect to the building. Undergrounding technology became more prevalent in the 1970s in the U.S., and it tends to be associated with larger, more valuable homes. For each home we query, we obtain the year built, the Zillow-estimated value, the finished size of the home, the parcel (lot) size, and the tax assessment as our covariates. The status of the power line nearest to the building (overhead vs. underground) is the response variable. We train random forest classifier and validate the model with holdout tests. We then finalize the model by training the random forest with that variable subset with the whole dataset.

In practice, due to the difficulties of scraping the housing data (Zillow API query is limited to 1000/day), we could only get house information for a small portion of houses. For example, it took us 14 days to retrieve information for 60,000 houses out of the 650,000 houses for Franklin County, Ohio. As a result, the line type of 90% of powerlines cannot be directly predicted by the model due to a lack of housing information on those roads. Therefore, we first predict the powerline type for each house with existing data. Then we aggregate each house's powerline type to its closest road by using the majority. For example, for a given synthetic powerline, if we scraped four houses along the powerline and three of the houses are predicted to be overhead and one to be underground, then we classify the powerline type of this synthetic powerline to be overhead. If there is a tie, we choose overhead as the powerline type.

Then for synthetic powerlines without any scraped houses along them, we spawn their powerline types from powerlines that have been classified. To do this, we iterate through all the unclassified powerlines, and find their connected powerlines. The powerline type of undetermined powerlines will be randomly sampled from their neighboring and classified powerlines' types. If none of an unclassified powerline has been determined, this powerline will remain undetermined until the next round. After all the unclassified powerlines have been iterated, we start the process again for any powerlines that are still unclassified. The process ends when all the powerlines are assigned a powerline type.

3.3. Power outage simulation

With the synthetic power system generated, we can use it to simulate power outages due to natural hazards through use of infrastructure fragility curves. The simulation scheme is shown in Fig. 5. This algorithm estimates the probability of each customer point losing power due to a given hazard loading scenario.

Our synthetic power system can be used to simulate power outages for many types of natural hazards provided we have (1) a proper spatial map of the loading due to the hazard and (2) valid fragility functions for poles and substations that converts the hazard loading parameters into asset-level failure probabilities. We explain the framework with using strong wind events as an example.

As a starting point, we have as our inputs the network layouts for each substation cluster, a map of wind speeds over the study area, and fragility functions that give the probability of failure as a function of wind speed. Network layouts consist of the distribution layouts for each substation cluster of the study area. In our example we use three-second wind gust speed as our hazard loading measure. The fragility functions then give the probability of pole failures for a given wind speed. For each substation cluster, we simulate the power outages with a sufficient number of replications for the probability of power outages of each customer to converge. Within the simulation, we first simulate which infrastructures fail, i.e., poles and substations. Then we change the network structure by removing these assets. The last step is to check if each customer still has connectivity to the substation on the damaged network. If there is no route to connect them to the substation because of pole or power line failures, or their substations are damaged, they will lose power. As a result, by integrating across all of the replications for each house, we can estimate the probability of a house losing power due to the simulated event, which can be informative to government, utility companies, and decision makers.

4. Results

4.1. Case study

Franklin County, Ohio is our primary case study. There are approximately 655,440 buildings in this county with a population of approximately 1.3 million. Our algorithm can conceptually be applied to any city, county, or an entire state if the required data is available and there are sufficient computational resources. We choose Franklin county, Ohio because we have the historical power outage data to compare our algorithm's output to, and we have the distribution system's actual layout which enables us to validate the layout generated by our model. We present our validations from two aspects, the similarity of our synthetic networks to the actual distribution network with multiple metrics, and the performance of our model in simulating a historical power outage event. We then show an application of the approach to a different area, Corpus Christi, Texas, for which we have some, but less spatially detailed, information to validate against. This provides at least some confidence in the ability to generalize the approach to other locations.

4.2. Network validation

The purpose of our synthetic model is to provide informative risk assessment of the impacts of hazards on the power system of an area. As such, comparing the end to end prediction performance using our model against power outages from historical events is important. However, we first compare network structure of our synthetic power network with the actual distribution network and evaluate how accurate our overhead/underground classifier is in comparison to the actual layout. These similarity-based validation metrics are important because, if the algorithm accurately reflects key properties of real systems, we can have more confidence in the applicability of the approach to hazards beyond the limited set of validation events that we have to compare to.

4.2.1. Network similarity

To compare the similarity between networks, we first compare global network parameters, such as the average nodal degree, betweenness centrality, number of circles, and the total length of the network. These parameters are commonly used to represent the attributes of network graphs, but they are not necessarily informative for risk assessment purposes [33]. We compare these metrics for each substation cluster we create with all three methods.

Nodal degree is the number of nodes that directly connect to a given node. From Table 1 we see that the average nodal degrees we generate with all three approaches are all close to the actual distribution network. This is clearly the result of the fundamental structure of distribution network because most of the nodes (poles) in the graph are connected with only two other nodes aiming to deliver electricity. Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. The average betweenness centrality is also at the same level between the two systems. The average Number of Circle describes the average number of loops in each

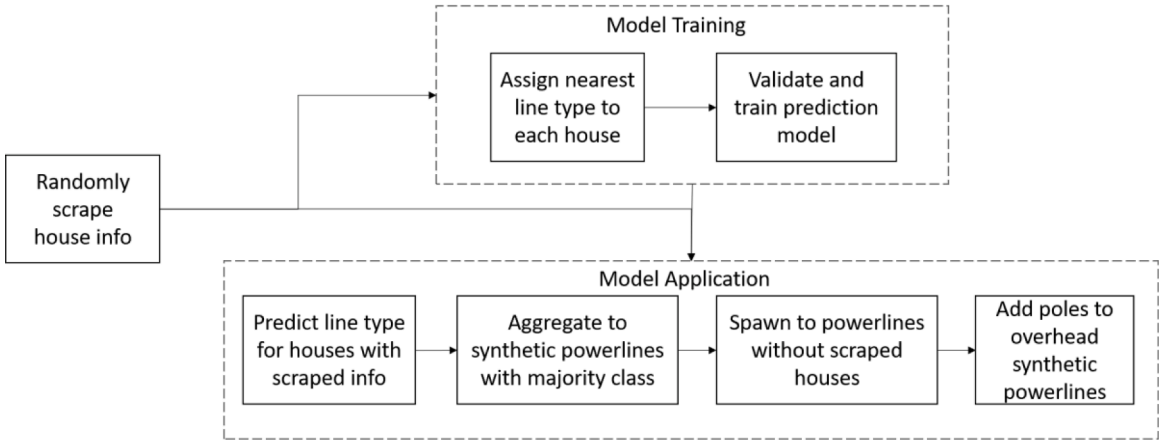


Fig. 4. Classification process for determining overhead vs underground status of each line.

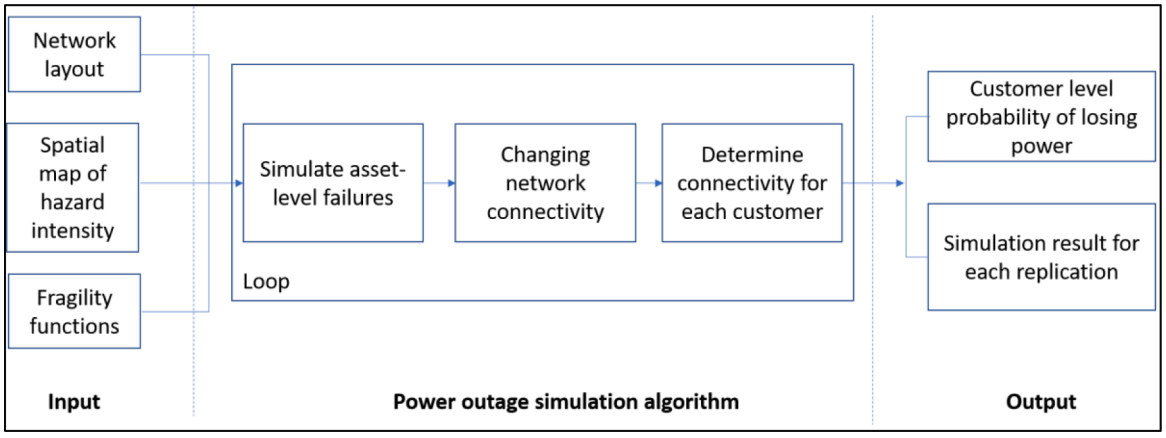


Fig 5. Simulation framework for estimating power outages at the household level given a network layout and asset-level fragility curves.

Table 1
Results of comparing network topology measures and network distance between the real and synthetic distribution systems.

	Average Nodal Degree	Average Betweenness Centrality	Average Number of Circle in Overhead	Total Length of Lines (Meters)	Mean Absolute Difference (m)	Pearson correlation
Actual System	2.066	0.0429	2.44	1.04×10^8	–	–
Steiner Tree Heuristic	1.999	0.0400	0	1.08×10^8	1426	0.699
Steiner Tree + Kmean	1.999	0.0389	0	1.09×10^8	882	0.745
Shortest Path	1.999	0.0363	0	1.07×10^8	693	0.842

substation cluster. We only consider the overhead system in this metric because underground systems are typically built as open loops with switches for extra robustness and from the data available to us form the utility, we cannot determine the actual connectivity of the undergrounding network. Table 1 shows that circles are rare in real overhead distribution networks which substantiates our assumption that the system is highly radial. Our synthetic systems are, by design, completely radial. Lastly, we also compare the total length of our synthetic system versus the actual system and we find they are very close. These results show that the approximation using road network to the distribution network is reasonable in terms of network topological parameters.

One of the most important metrics for risk assessment purposes is the network distance from each customer to its substation. Due to the nature of radial systems, each customer is served by one substation. The probability of losing power for a customer is positively correlated to its network distance to the substation because the further the distance, the higher the number of potential failure points between the customer and the substation.

We compare the actual customer to substation network distance from the distribution layout with the synthetic networks generated by the three algorithms we propose. For each customer, on the synthetic network, we calculate the shortest path to the substation it is assigned to. If a customer is within multiple substation clusters, we use the shortest one for comparison. The result is shown in the last two columns in Table 1. In terms of each buildings distance to the nearest substation, the best model we find is the shortest path model. The average absolute difference in distance to the actual network is 693 m. We also calculate the Pearson correlation and the shortest path model outperforms the other two models as well. Pearson correlation coefficient measures the linear correlation between two variables. These results suggest that our model, while not perfect, generates synthetic systems with customer distances to substations that correlate well with the actual values.

4.2.2. Overhead/underground power lines classification validation

The second step is to evaluate the accuracy of our overhead/underground classifier. We first test the accuracy of the random forest model given our dataset. The in-sample prediction accuracy with the whole dataset is 100%. More meaningful is the out of sample accuracy. The average out-of-sample prediction accuracy in 30 repeated random holdouts is 91%. 10.6% of the synthetic powerlines are directly predicted from the dataset and the accuracy is 100% (due to in-sample prediction). After applying the powerline type spawning algorithm, the overall prediction accuracy for the whole network is 84.1% for our study region. One issue of our model is that it cannot capture commercial buildings because Zillow does not provide such information. One potential solution is to identify commercial area and use a pre-defined line type for power lines surrounding commercial areas.

The level of accuracy of our model is strong even given our limited dataset, and it may improve if information on more buildings or other sources of additional real estate data become available. We have also directly compared the generated systems to the actual systems visually on a map. However, for reasons of security, the data provider does not allow the actual maps to be shown. However, the systems are visually similar.

4.3. Extreme weather simulation

4.3.1. Franklin county, Ohio - Derecho, 2012

In this section, we use the model we developed to predict power outages the Derecho that impacted Franklin County, Ohio on June 29, 2012. The event caused more than 50% of customers in the county to lose power. We have the utility outage data, aggregated to 5 km grid cells, to compare with our model outputs. During the event, wind was the driving force of power outages. We retrieved the maximum gust wind speed during the event for all the airports in or near the county and interpolate to get the gust wind speed for each pole given the pole's spatial location to those airports.

We applied the fragility curves developed by Darestani and Shafieezade [31]. The fragility of poles under wind events is determined by the pole class, wind speed and direction, age, diameter of conductors, heights. By inputting these parameters, we can estimate the probabilities of pole failure and use these to simulate the change of network connectivity. We repeated out analysis for different combinations of pole classes and ages and conduct sensitivity analysis to cover the gap of unknown pole information. We do not consider damage to falling trees due to lack of information, though we acknowledge that tree failures can be substantial causes of outages. We conduct a convergence test on the average customers without power for several substation clusters with 50,000 replications. We find that the relative difference for the average customers without power is less than 1% after 10,000 replications.

We compare our simulation results with the actual event in 5 km by 5 km grids which are shown in Fig. 6. We assumed all the poles are the same throughout the system in each run, and we tested different types of poles in each scenario: 60-year-old class 4 poles, 50-year-old class 5 poles, and 60-year-old class 5 poles. These pole ages were chosen based on the average age of buildings in the region. The class of poles is determined by the minimum circumference that depends on the species of tree and the length of the pole [46]. Higher-class poles (e.g. class five) typically can hold less horizontal load than lower-class poles (e.g. class four). Class four and five poles are typically used in distribution systems in the U.S.

From these simulation scenarios we gain insights into the relative risks of losing power. In all cases (plots (b)-(d) in Fig. 6), the model is

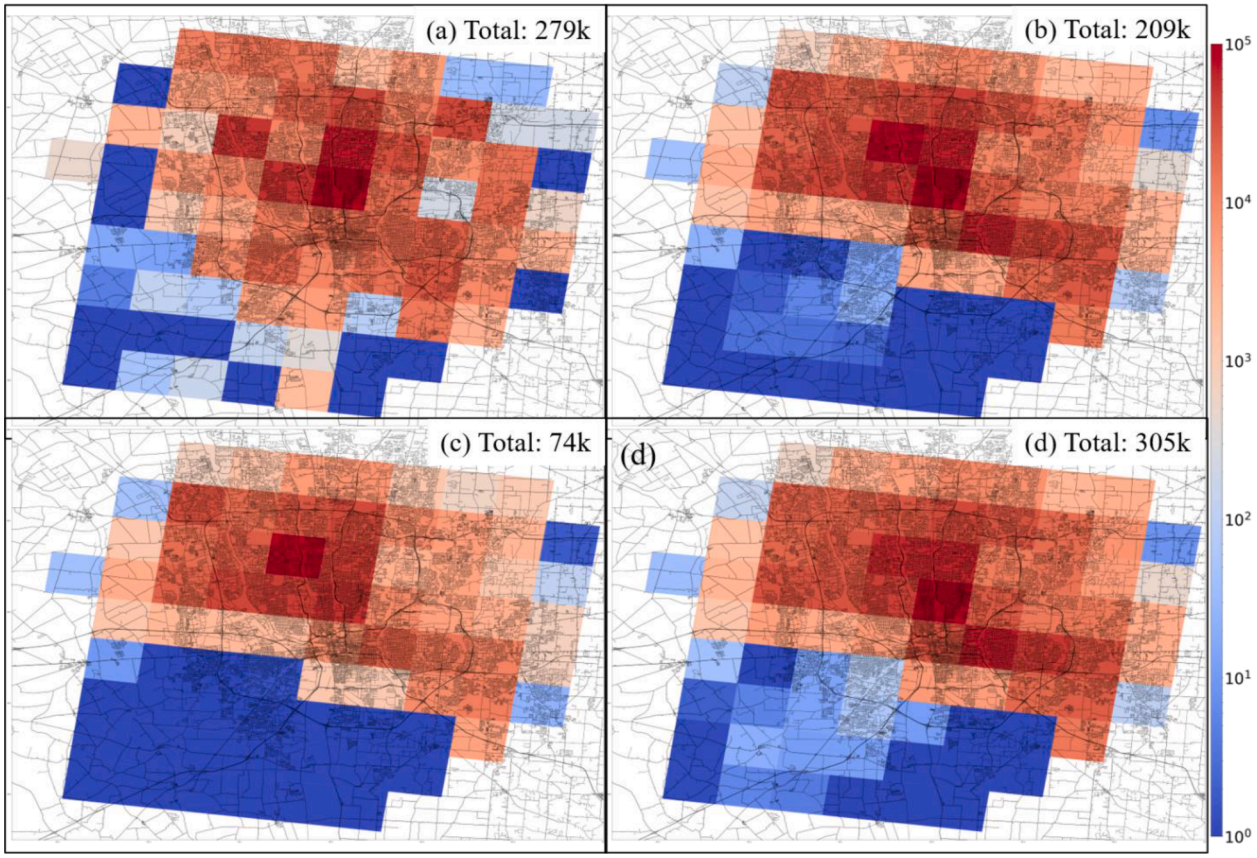


Fig. 6. Comparison of actual and simulated power outages for the 2012 Derecho in Franklin County, OH. The legend shows the average number of customers without power in each grid cell. (a) actual power outage during the derecho, (b) If all the poles are class 4 and age 60, (c) If all the poles are class 5 and age 50, (d) If all the poles are class 5 and age 60. Total number of customers without power is also shown in each scenario.

capturing the power outages on the northern side of the region reasonably well but underestimating outages in the southern portions of the region. Part of the reason for this is because we do not have the most accurate wind speed data for all grid cells, but only the seven major airports in this area. These airports are mostly located in the northern part of the county. Plots (b) and (d) are both for 60-year-old poles. These two model runs give a total number of customers without power relatively close to the actual 279,000. The major difficulty in simulating this Derecho event is the to a lack of information on the hazard loading (i.e., wind speeds). We use only gust wind speed at major airports to approximate the gust wind speeds throughout the system. Informal conversations with the meteorologists from the utility that provided the data revealed that wind speeds were highly spatially variable during this event, and we lack the data to capture these local differences. To improve the simulation results, a more detailed wind speed map would be helpful.

4.3.2. Corpus Christi - Harvey, 2017

In order to test the generalizability of the approach to other areas and other events, we also simulated the impacts of Hurricane Harvey in southern Texas. We do not have the actual outage data for this area in the same level of detail but have obtained estimates of outages from media reports. We also do not have the full system topology to compare to. We should note that we are focusing on the area of Texas that experienced Harvey as strong wind event, not primarily a flooding event (i.e., we are not considering the Houston area).

Hurricane Harvey made landfall in Texas on August 23, 2017. The hurricane caused considerable damage loss of life. During the event, widespread power outages occurred in multiple major cities due to strong wind, hurricane surge, and flooding caused by rainfall. We selected Corpus Christi to test our model because the majority of power outages there were wind driven. We use the best track estimation[4] of hurricane Harvey and an existing hurricane wind field model [47] to estimate the 3-second gust wind speed for each distribution substation. We then apply the same wind speed from each substation to all the poles connected to that substation and simulated outages. The results are shown in Fig. 7. Red areas are those areas where the buildings are more likely to lose power and blue are less likely to lose power. We assume the parameter for poles to be 40-year-old class 4 poles, 50-year-old class 4 poles, and 40-year-old class 5 poles based on the building stock age for the city.

As a comparison, from media reports, the peak number of customers without power reported by AEP, the utility serving the area, during Harvey for Corpus Christi was approximately 91,500. The majority of the outages were in the Midtown area and the Southside area while the Northwest area was mildly damaged. With our simulation, we estimate the average number of customers without power as 67,000, 84,000, and 82,000, where the estimates from the latter two sets of pole parameters are close to the actual value. The model estimates the risk of losing power to be high in the midtown area for all three sets of pole parameters. For the southside area, the outage estimates are lower because the model estimates that the three substations in this area have primarily underground power lines serving customers. The other substations are severely damaged. For the northwest area, especially in the scenario of 40-year-old class 4 poles, there are less customers damaged. While we do not have detailed ground truth data to compare to, overall the model is in agreement with the media reports of outages, at least at a high level. Overall, these simulation results can be a good source of information to the public, critical infrastructures, and utility companies to assess the potential chances of losing power and find a better way to mitigate.

[4] https://www.nhc.noaa.gov/gis/archive_besttrack_results.php?id=al09&year=2017&name=Hurricane%20HARVEY.

5. Conclusion

We develop a model to use synthetic network generation to cover a critical gap in power outage risk assessment research. We use publicly available data to create a synthetic version of the distribution system given building information, substation locations, and road networks. The method is generalizable to larger scales, from city level to county level, and potentially to state level. We tested our approach for Franklin county, Ohio, and Corpus Christi, TX. For Franklin county, we compared our network with the actual distribution network with multiple measures and found that the model approximated the actual system well in terms of topological characteristics. We also simulated two major power outage events, the Derecho in Ohio in 2012 and hurricane Harvey in Corpus Christi in 2017. The results of these simulations show that the model can produce accurate estimates of power outages provided accurate hazard loading maps are available.

However, there are several limitations in this method that needs to be addressed. What we are creating is a synthetic distribution network that can be representative of the real-life system and can be used to estimate damage to the distribution system and household likelihood of losing power in the US. For other countries, it is not clear at this point if the method is applicable because 1) certain assumptions may no longer be held (i.e., that the distribution system is radial), 2) certain input data may become harder to acquire, (i.e., building information and substation locations), and 3) the overhead and underground power line classifier may no longer work because of differences in how the system is designed. Some of these limitations can be resolved by introducing extra models to impute missing information, such as using population and power consumption information to create synthetic distribution substations. In the meantime, with better input data, the performance of certain models can be improved. For example, with a more detailed building information, the classification of lines as overhead or underground may become more accurate. In addition, more information regarding the age of certain infrastructure such as poles can help improve the outage estimation. Overall, our work shows the effectiveness of the method with limited information.

Another major challenge of the proposed method is its scalability. While the method has been designed for use at the scale of a metropolitan area, there may be interest in scaling up to a state or national level. Scaling up to state-wide or nation-wide analysis with this approach would have three major challenges: data collection, computing power limitations, and fragility function limitations. The first challenge would be how to collect the necessary input data for the model if applying the model to the state or national level. Road locations, building locations, and substation locations can all be retrieved publicly for the entire U.S., but the effort required to gather this data grows substantially for larger application domains. For building information, one possibility would be to switch to a commercially available database of building locations and characteristics. The second challenge is the computational power required for both network generation and network simulation. Highly optimized code and parallelization can increase the feasibility of this application. Variance reduction and sampling strategies can be leveraged for faster convergence. Even with these strategies, computational effort will remain a challenge for very large spatial domains. The last challenge is regarding fragility functions. A comprehensive set of fragility functions for vulnerable components of the distribution system for the hazard type being modeled is crucial in improving the accuracy of the estimation. For certain components under some disasters, the fragility functions available may not exist up to date. Resolving this issue will rely on the development of improved fragility functions.

Though only tested for wind events in this paper, we have run the model for other hazard types. We use only wind events as an example in this paper because of the availability of data to validate the performance of our model against historical data. Using the model for a different hazard requires hazard-appropriate fragility functions and potentially

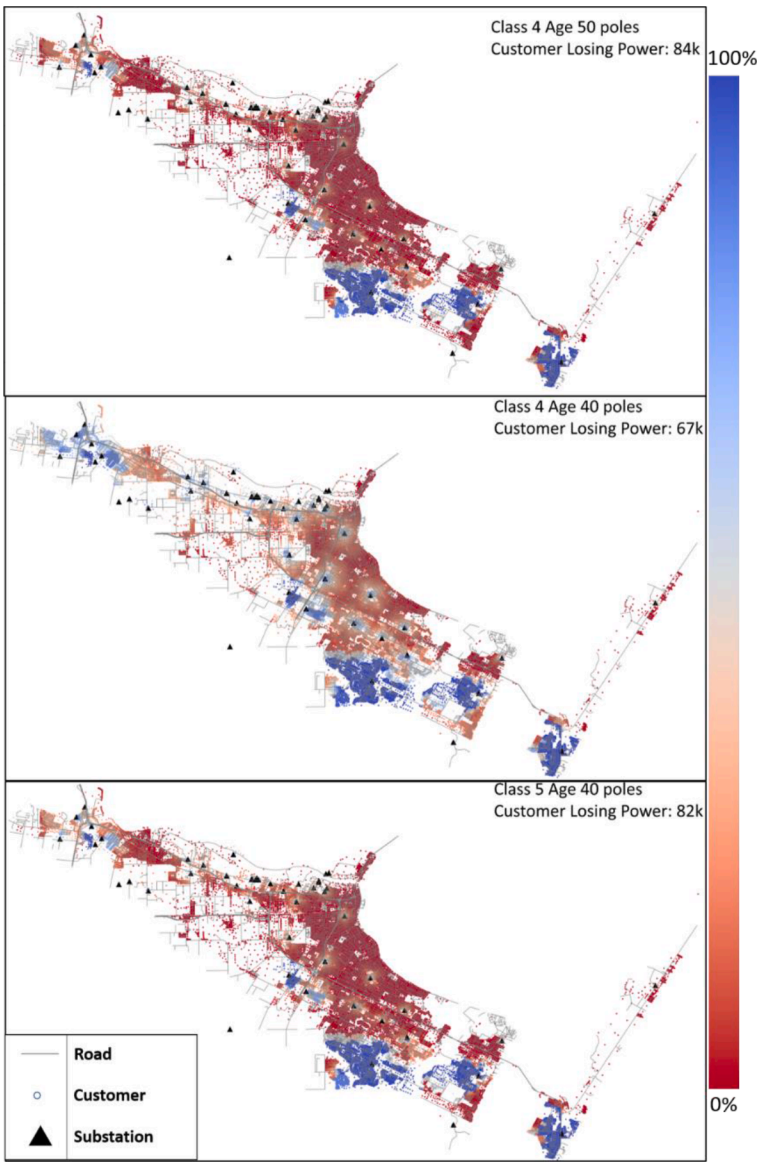


Fig. 7. Simulation for Hurricane Harvey, 2017 for Corpus Christi. The legend shows the probability each customer with power. Closing to 100% indicates the customer will have a high probability of having power. Closing to 0% indicates the customer will have a lower probability of having power.

the inclusion of additional system components if they are vulnerable under the other hazards. For example, for earthquakes, the damage to the distribution system comes largely from damage to low-voltage substations and utility poles, requiring seismic fragility functions for these components. On the other hand, for flooding, the height of certain utility equipment, such as substation transformers should be explicitly modelled to estimate their failure probabilities.

Overall, our model provides a novel approach for estimating the building-level probability of losing power for a natural hazard event. We have shown that this approach can yield accurate predictions provided an accurate map of hazard loading is used and appropriate fragility functions are used. This approach has the potentially to both improve predictive accuracy of power outage estimation and to provide substantially finer spatial detail in predictions than existing approaches provide.

CRediT authorship contribution statement

Chengwei Zhai: Conceptualization, Methodology, Software, Formal analysis, Validation, Investigation, Data curation, Writing - original draft, Visualization. **Thomas Ying-jeh Chen:** Writing - review & editing, Conceptualization, Visualization, Investigation, Formal analysis, Software. **Anna Grace White:** Writing - review & editing, Visualization, Investigation, Conceptualization. **Seth David Guikema:** Writing - review & editing, Conceptualization, Methodology, Supervision, Funding acquisition, Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was funded by One Concern, Inc. through a grant to the University of Michigan. One of the authors (Guikema) also has an ownership stake in One Concern, but independence of the research was managed through University of Michigan procedures. The opinions in this paper are those of the authors and do not necessarily reflect those of

the sponsor.

References

[1] Hines P, Apt J, Talukdar S. Trends in the history of large blackouts in the United States. In: IEEE power and energy society 2008 general meeting: conversion and delivery of electrical energy in the 21st century. PES; 2008.

[2] Kezunovic M, Zoran O, Tatjana D, Bei Zhang JS, Payman D, Po-Chen C. Predicting spatiotemporal impacts of weather on power systems using big data science. In: Data Science and Big Data: An Environment of Computational Intelligence. 24. Cham: Springer; 2017. p. 265–99.

[3] Eisenbach Consulting LLC. 9 of the worst power outages in United States history. Electric Choice 2017 [Online]. Available, https://www.electricchoice.com/blog/worst-power-outages-in-united-states-history/.

[4] B.W. Johnson, "After the disaster: utility restoration cost recovery exhibit MP1," 2005.

[5] Guikema SD, Nateghi R, Quiring SM, Staid A, Reilly AC, Gao M. Predicting hurricane power outages to support storm response planning. IEEE Access 2014;2:1364–73.

[6] Shafieezadeh A, Onyewuchi UP, Begovic MM, Desroches R. Age-dependent fragility models of utility wood poles in power distribution networks against extreme wind hazards. IEEE Trans Power Deliv 2014;29(1):131–9.

[7] Salman AM, Li Y, Stewart MG. Evaluating system reliability and targeted hardening strategies of power distribution systems subjected to hurricanes. Reliab Eng Syst Saf 2015;144(Dec.):319–33.

[8] R.J. Campbell, "CRS report for congress weather-related power outages and electric system resiliency specialist in energy policy weather-related power outages and electric system resiliency congressional research service weather-related power outages and electric system resiliency congressional research service," 2012.

[9] Davidson RA, Liu H, Sarpong K, Sparks P, Rosowsky DV. Electric power distribution system performance in Carolina hurricanes. Nat Hazards Rev 2003;4(Feb. (1)):36–45.

[10] Deka D, Backhaus S, Chertkov M. Learning topology of distribution grids using only terminal node measurements. In: 2016 IEEE international conference on smart grid communications, smart grid comm 2016; 2016. p. 205–11.

[11] Abdul Rahman F, Varuttamaseni A, Kintner-Meyer M, Lee JC. Application of fault tree analysis for customer reliability assessment of a distribution power system. Reliab Eng Syst Saf 2013;111(Mar.):76–85.

[12] Davidson RA, Liu H, Sarpong K, Sparks P, Rosowsky DV. Electric power distribution system performance in Carolina hurricanes. Nat Hazards Rev 2003;4(Feb. (1)):36–45.

[13] Han SR, Guikema SD, Quiring SM. Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models. Risk Anal 2009;29(Oct. (10)):1443–53.

[14] Han SR, Guikema SD, Quiring SM, Lee KH, Rosowsky D, Davidson RA. Estimating the spatial distribution of power outages during hurricanes in the Gulf coast region. Reliab Eng Syst Saf 2009;94(2):199–210.

[15] Guikema SD, Quiring SM, Han SR. Prestorm estimation of hurricane damage to electric power distribution systems. Risk Anal 2010;30(12):1744–52.

[16] Nateghi R, Guikema S, Quiring SM. Power outage estimation for tropical cyclones: improved accuracy with simpler models. Risk Anal. 2014;34(6):1069–78.

[17] McRoberts DB, Quiring SM, Guikema SD. Improving hurricane power outage prediction models through the inclusion of local environmental factors. Risk Anal. 2018;38(12):2722–37.

[18] Shashaani S, Guikema SD, Zhai C, Pino JV, Quiring SM. Multi-stage prediction for zero-inflated hurricane induced power outages. IEEE Access 2018;6:62432–49.

[19] Kabir E, Guikema SD, Quiring SM. Predicting thunderstorm-induced power outages to support utility restoration. IEEE Trans Power Syst 2019;34(Nov. (6)):4370–81.

[20] Han SR, Guikema SD, Quiring SM, Lee KH, Rosowsky D, Davidson RA. Estimating the spatial distribution of power outages during hurricanes in the Gulf coast region. Reliab Eng Syst Saf 2009;94(2):199–210.

[21] He J, Wanik DW, Hartman BM, Anagnostou EN, Astitha M, Frediani MEB. Nonparametric tree-based predictive modeling of storm outages on an electric distribution network. Risk Anal 2017;37(Mar. (3)):441–58.

[22] Guikema SD. Natural disaster risk analysis for critical infrastructure systems: an approach based on statistical learning theory. Reliab Eng Syst Saf 2009;94(Apr. (4)):855–60.

[23] Guikema SD, Quiring SM. Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data. Reliab Eng Syst Saf 2012;99:178–82.

[24] Cerrai D, et al. Predicting storm outages through new representations of weather and vegetation. IEEE Access 2019;7:29639–54.

[25] D'Amico DF, Quiring SM, Maderia CM, McRoberts DB. Improving the hurricane outage prediction model by including tree species. Clim Risk Manag 2019;25(Jan.):100193.

[26] Staid A, Guikema SD, Nateghi R, Quiring SM, Gao MZ. Simulation of tropical cyclone impacts to the U.S. power system under climate change scenarios. Clim Change 2014;127(3–4):535–46.

[27] Vickery PJ, Skerlj PF, Lin J, Twisdale LA, Young MA, Lavelle FM. HAZUS-MH hurricane model methodology. II: damage and loss estimation. Nat Hazards Rev 2006;7(May (2)):94–103.

[28] Kircher CA, Whitman RV, Holmes WT. HAZUS earthquake loss estimation methods. Nat Hazards Rev 2006;7(May (2)):45–59.

[29] Vanzi I. Seismic reliability of electric power networks: methodology and application. Struct Saf 1996;18(4):311–27.

[30] Han S-R, Rosowsky D, Guikema S. Integrating models and data to estimate the structural reliability of utility poles during hurricanes. Risk Anal 2014;34(Jun. (6)):1079–94.

[31] Mohammadi Darestani Y, Shafieezadeh A. Multi-dimensional wind fragility functions for wood utility poles. Eng Struct. 2019;183(Mar.):937–48.

[32] Guikema SD, Davidson RA, Liu H. Statistical models of the effects of tree trimming on power system outages. IEEE Trans Power Deliv 2006;21(3):1549–57.

[33] S. Larocca, J. Johansson, H. Hassel, and S. Guikema, "Topological performance measures as surrogates for physical flow models for risk and vulnerability analysis for electric power systems," 2013.

[34] Birchfield AB, Gegner KM, Xu T, Shetye KS, Overbye TJ. Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies. IEEE Trans Power Syst 2017;32(2):1502–10.

[35] Pahwa S, Scoglio C, Scala A. Abruptness of cascade failures in power grids. Sci Rep 2014;4.

[36] Schultz P, Heitzig J, Kurths J. A random growth model for power grids and other spatially embedded infrastructure networks. Eur Phys J Spec Top 2014;223(Oct. 12):2593–610.

[37] Soltan S, Zussman G. Generation of synthetic spatially embedded power grid networks. In: IEEE power and energy society general meeting. 2016; 2016. Novem.

[38] Pisano G, et al. Synthetic models of distribution networks based on open data and georeferenced information. Energies 2019;12(Nov. (23)):4500.

[39] Schweitzer E, Member S, Scaglione A, Monti A, Member S, Andrea Pagani G. Automated generation algorithm for synthetic medium voltage radial distribution systems. IEEE J Emerg Sel Top Circ Syst 2017;7(2).

[40] Miranda V, Ranito JV, Proena LM. Genetic algorithms in optimal multistage distribution network planning. IEEE Trans Power Syst 1994;9(4):1927–33.

[41] Valenzuela A, Inga E, Simani S. Planning of a resilient underground distribution network using georeferenced data. Energies Feb. 2019;12(4):644.

[42] Yuan W, Wang J, Qiu F, Chen C, Kang C, Zeng B. Robust optimization-based resilient distribution network planning against natural disasters. IEEE Trans Smart Grid 2016;7(Nov. (6)):2817–26.

[43] N. A. of E. and N. R. C. National Academy of Sciences. America's energy future: technology and transformation. National Academies Press; 2010.

[44] H. Takahashi and A. Matsuyama, "An approximate solution for the Steiner tree problem in graphs." 1980.

[45] Garey MR, Johnson DS. The rectilinear Steiner tree problem is NP-complete. SIAM J Appl Math 1977;32(Jun. (4)):826–34.

[46] R. Wolfe, "Standard specifications for wood poles," 1997.

[47] Holland GJ, Belanger JI, Fritz A. A revised model for radial profiles of hurricane winds. Mon Weather Rev 2010;138(Dec. (12)):4393–401.