Reliability Engineering and System Safety 212 (2021) 107621

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Understanding key factors affecting power systems resilience

Lijuan Shen [a], Yanlin Tang [b,*], Loon Ching Tang [c,*]

[a] ETH Zurich, Future Resilient Systems, Singapore-ETH Centre, 138602, Singapore
[b] Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE, School of Statistics, East China Normal University, Shanghai, 200062, China
[c] Department of Industrial Systems Engineering and Management, National University of Singapore, 117576, Singapore

ARTICLE INFO

Keywords:
Feature selection
Power systems
Severe weather
Performance loss
Recovery time

ABSTRACT

In this paper, we study the key factors that impact on power systems resilience under severe weather-induced disruptions from three dimensions: the extrinsic disruptions, the intrinsic capacities of a system and the effectiveness of recovery. Using 12 years of historical blackout data from 2007 to 2018 in the U.S., we apply various group selection and bi-level selection methods to identify the key predictor groups as well as factors within-group that affect power system resilience. After deleting the predictors which are fully or highly correlated with others, we split the remaining 39 candidate predictors into 8 natural groups and consider the number of customers affected and the recovery time as response variables. To ensure stability of the selection process, we adopt the random subsampling method to rank the importance of the groups and key predictors. It is found that the disruption types from the extrinsic disruptions dimension have a significant impact on the resilience of power systems, especially for the hurricanes with high scales. From the intrinsic capabilities dimension, the demographic group has a large impact on the number of customers affected. The number of customers affected tends to be large in highly urbanized areas with large population. From the effectiveness of recovery dimension, the group of economics is top selected for the recovery time. It is found that the power system tends to be more resilient with a better economic health. Feature selection under quantile regression is also conducted as the histograms show that the distributions of the responses are skewed and heavy-tailed. It is found that the recovery time is also greatly affected by the investment on the compliance and enforcement program from the North American Electric Reliability Corporation. In summary, our analysis provides interesting insights for understanding power system resilience and developing strategies to enhance the resilience.

1. Introduction

It has been found that severe weather conditions, such as hurricanes, thunderstorms and winter storms are the leading cause of blackouts in the U.S. [1]. With the increasing frequency and intensity of severe weather due to climate change, severe weather-induced power outages can affect millions of people and result in billions of dollars of economic losses. The effect of severe weather is becoming more damaging on power systems due to urbanization and economic developments that lead to increased electricity demand and greater scale and interdependencies between urban systems. For example, Hurricane Sandy in 2012 resulted in power loss of 8 million people with estimated damage up to $70 billion; Hurricane Irene in 2011 resulted in power loss of 6 million people with estimated damage up to $10 billion. Therefore, it is crucial to understand how the resilience of power systems induced by severe weather is affected by various factors.

Resilience is commonly defined as the ability of a system in mitigating the effects of disruptions and to recover quickly from disruptions. Bruneau et al. [2] argued that the main features of resilience are robustness, redundancy, resourcefulness, and rapidity. Vugrin et al. [3] proposed a framework using three system capacities to formulate how inherent properties of a system could determine system resilience, which are absorptive capacity, adaptive capacity, and restorative capacity. Shen et al. [4] defined resilience as a function of intrinsic capacities of a system, the effectiveness of recovery and the extrinsic random shock process. Bruneau et al. [2] introduced the resilience triangle paradigm to measure seismic resilience of communities, where the triangle is interpreted as the expected performance loss over time. Henry and Ramirez-Marquez [5] quantified resilience as the ratio of recovery to loss, which is time-dependent. In addition, some studies integrated hazard models, component fragility models and restoration models

* Corresponding authors.
E-mail addresses: yltang@fem.ecnu.edu.cn (Y. Tang), isetlc@nus.edu.sg (L.C. Tang).

https://doi.org/10.1016/j.ress.2021.107621
Received 5 July 2020; Received in revised form 15 December 2020; Accepted 11 March 2021
Available online 26 March 2021
0951-8320/© 2021 Elsevier Ltd. All rights reserved.

to measure resilience [6,7]. A comprehensive review on resilience definitions and measures can be found in [8] and [4].

A resilient power system should be able to ensure sustained service deliveries when disrupted from a disruption either by mitigating its effect or by speedy recovery or both. Much research effort has been devoted to quantifying the resilience of power systems to extreme weather. For example, Shen et al. [1] assessed the trend in resilience of power systems, looking into the time between disruptions, the performance loss of each disruption and the time needed for recovery. Figueroa-Candia et al. [9] proposed a modeling framework based on resilience for the evaluation and optimization of restoration policies for electric power distribution systems subject to extreme weather events. Please refer to Shen et al. [4] for a detailed review of resilience measures. Some research work has been developed to forecast the power outage durations. For example, Liu et al. [10] proposed a statistical forecasting method for electric power restoration times in hurricanes and ice storms, which is useful to plan and optimize the restoration efforts. Nateghi et al. [11] compared statistical methods for modeling power outage durations during hurricanes and examined the predictive accuracy of these methods. Wang et al. [12] reviewed some forecast models for power systems damages and outage durations, including both the statistical models and simulation based models. Mukherjee et al. [13] developed a hybrid support vector machine-random forest predictive framework to characterize the key predictors of severe weather-induced sustained power outages.

Although much research effort has been focused on the modeling and quantification of power system resilience, scant attention has been paid to identifying the key factors affecting resilience quantitatively and systematically. In addition, much literature focus on modeling power systems damages and outage durations under a specific type of hazard, such as hurricanes, and there lacks studies on multi-hazards. In this paper, we look into the power distribution system and present a statistical model to relate resilience to various factors so that resilience under different types of weather induced hazards could be assessed quantitatively. The results can provide fundamental insights on understanding power systems resilience and developing strategies to enhance resilience.

Following [4], resilience can be viewed as a function of extrinsic disruptions, intrinsic capacities of a system and the effectiveness of recovery. By looking into the three dimensions, the candidate predictors analyzed in this study are naturally grouped. For extrinsic disruptions, we discuss the types and scales of disruptions and the climate characteristics. Under intrinsic capacities, we consider the electricity price, the electricity consumption, the demographic characteristics and the geographic characteristics. For the effectiveness of recovery, some economic characteristics and technical investments are discussed. As the candidate predictors present a grouping structure, we are interested in selecting both important groups and factors within-group, which is defined as a bi-level selection problem. The historical data from 2007 to 2018 in the U.S. are used in the study. Please note that the predictors analyzed in this study could be extended further if more data are available.

Feature selection is an important topic in regression analysis, theoretically and practically. Generally, to attenuate possible modeling biases, a large number of potential predictors are included at the initial stage of data analysis. With proper feature selection methods, a more interpretable model could be built to avoid over-fitting in prediction and estimation. Furthermore, predictors can often be grouped in many applications, yet the individual predictor may also be of interest. The conventional analysis of variance (ANOVA) method selects significant factors sequentially and cannot be used for bi-level selection. Penalized approaches for feature selection have gained popularity in recent years, which do variable selection and estimation simultaneously in a one-shot and is more preferable than ANOVA. Penalties designed for individual feature selection include the least absolute shrinkage and selection operator [14, LASSO], smoothly clipped absolute deviation penalty [15, SCAD] and minimax concave penalty [16, MCP], etc. By applying these penalties to group selection and bi-level selection, some methods have been developed. A review of penalized regression methods for both group selection and bi-level selection could be found in [17].

In this study, we apply both group selection and bi-level selection methods to identify key groups as well as factors within-group that impact on power systems resilience. Random subsampling, a robust algorithm, is combined with the feature selection methods to rank the importance of the key predictors. More specifically, we look into two key components of power systems resilience: the absorptive capability and the recovery capability. The performance loss in terms of service deliveries depends on the shock magnitude as well as the system's ability to absorb the shock, which could be used to represent the absorptive capability. The recovery time depends on the level of the performance loss as well as the resources available for the recovery, such as the maintenance effort and emergency routines, which could be used to represent the recovery capability. In this paper, the number of customers affected and the recovery time are considered as candidate response variables.

The rest of the paper is organized as follows. In Section 2, the power grid data are introduced from three dimensions: the extrinsic disruptions, the intrinsic capabilities and the effectiveness of recovery. In Section 3, the candidate predictors and response variables are defined. The group selection and bi-level selection methods are reviewed. In Section 4, we apply the feature selection methods to the historical data of the U.S. power systems and the selection results are discussed. In Section 5, the results of feature selection under quantile regression are discussed as the distributions of the response variables tend to have heavy tails. Section 6 provides conclusions and directions for future research.

2. Power grid data

The U.S. Department of Energy (DOE) collects blackout information of power systems through the Electric Emergency Incident and Disturbance Report (Form OE-417). The blackout data includes the start time of disruptions, the restoration time, the affected state, the affected North American Electric Reliability Corporation (NERC) regions, the disruption types and the number of customers affected. The disruption types include winter storms, hurricanes, thunderstorm, transmission equipment failure, fuel supply emergency, voltage reduction, physical attack, etc. Since severe weather is the leading cause of blackouts in the U.S. [1], our analysis focuses on blackouts due to severe weather-induced events, looking into the blackouts data from 2007 to 2018.

As defined in [4], resilience is determined by the key drivers from the three dimensions: the extrinsic disruptions, the intrinsic capabilities and the effectiveness of recovery. For the disruption dimension, the disruption frequency and the disruption intensity need to be considered. The intrinsic capabilities are the abilities of a system to resist and absorb the disruption to mitigate its effects. The effectiveness of recovery determines the rate of restoration in system performance, representing the recovery speed and recovery time. It may depend on investments, recovery resources, and so on. Please refer to Table 7 in Appendix A for an overview of the full predictors from the three dimensions for analyzing the power system resilience. Some data are extracted from [18].

For the extrinsic disruptions, we first consider the disruption types. The disruption types of severe weather is classified into hurricanes, windstorms, thunderstorms, winter storms, other storms, wildfires, and others and unknown. Other storms indicate storms which do not belong to hurricanes, windstorms, thunderstorms and winter storms, e.g., hailstorm or mixed storms. Others and unknown include lightning, flooding, others types and not recorded types of severe weather. The disruption types of severe weather in each state of the U.S. from 2007 to 2018 is shown in Fig. 1. It shows that, (i) overall, the eastern

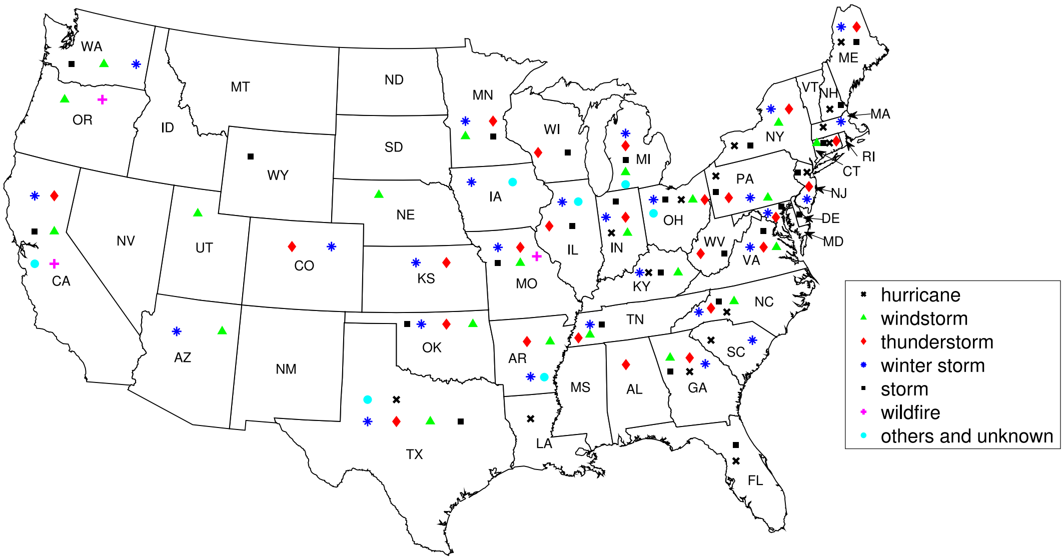

Fig. 1. The types of severe weather in the U.S. from 2007 to 2018.

U.S. was more prone to severe weather disruptions; (ii) for single state, Texas (TX), Ohio (OH) and California (CA) experienced more disruption types among all the states; (iii) the thunderstorms and winter storms were more common types among all the disruption types. The frequency of severe weather disruptions in each state is shown in Fig. 2. The figure shows that, (i) Michigan (MI) experienced more frequent disruptions, followed by Texas (TX) and California (CA); (ii) Texas (TX), Florida (FL), New York (NY) and Louisiana (LA) were exposed to hurricanes more frequently; (iii) Michigan (MI) experienced more frequent thunderstorms and windstorms; (iv) California (CA) experienced more frequent wildfires and winter storms.

Furthermore, the scale of hurricanes is considered to discuss the effect of disruptions intensity. The Saffir–Simpson hurricane wind scale is usually used to estimate potential property damage, which is a 1 to 5 rating based on a hurricane's sustained wind speed. Hurricanes reaching Category 3 and higher are considered severe hurricanes which may result in significant loss of life and damage. We classify hurricanes into two types in this work, where type I includes Category 1 and Category 2 hurricanes, and type II includes Category 3 and higher. It is found that the number of customers affected tends to be larger and the recovery time is longer when subjected to the type-II-hurricane, seeing the box-plots in Fig. 3. The result is consistent with the Saffir–Simpson hurricane wind scale, which indicates that Category 3 and higher hurricanes tend to have significant loss that may affect a large number of customers and result in prolonged power outages.

In addition, the climate characteristics are also discussed. We look into the Oceanic Niño Index (ONI), which is one of the primary indices used to monitor the El Nino-Southern Oscillation (ENSO). Based on a threshold of ± 0.5 °C for the ONI, the warm and cold periods could be defined.

For the intrinsic capacities of power systems, the factors considered include the electricity price, the electricity consumption, the demographic characteristics and the geographic characteristics. (i) The electricity price provides information about the three types of electricity price (residential, commercial and industrial) and the average electricity price for each state. (ii) The electricity consumption provides information about the total electricity consumption, the three types of electricity consumption (residential, commercial and industrial), and their corresponding percentages in each state. (iii) The demographic information includes the annual population for each state, the population percentages in urban areas and urban clusters, the population density of the urban areas, the urban clusters, and the rural areas; the number of total customers served, the numbers of the three types of customers served (residential, commercial and industrial), and their corresponding percentages in each state. (iv) The geographic information includes the percentages of the land areas of the urban, the urban clusters and the state, the percentages of the water area and inland water area for each state. (v) In addition, the eight climate regions in the U.S. are also considered, including Central, East North Central, Northeast, Northwest, South, Southeast, Southwest, West, and West North Central. Some regions contain few observations, as a consequence of high granularity in the region classification. For better illustration, we redefine the regions according to national weather service regions [19], dividing the U.S. into four climate regions: western, central, southern and eastern, as shown in Fig. 4.

For the effectiveness of recovery, some economic characteristics and types of technical investments are potential factors to be considered. The economic information includes the per capita real gross state product (GSP) of each state, the total per capita real gross domestic products (GDP) of the U.S., the percentage change of per capita real GSP from the previous year, the real GSP contributed by utility industry, the real GSP contributed by all industries, the ratio of the state utility's income to the U.S. utility's income, and the utility's contribution to the total GSP in each state. For technical investments, we consider investments in NERC programs which include (i) the reliability standards program; (ii) the compliance and enforcement program; (iii) the reliability assessment and performance analysis program; (iv) the training and education program; and (v) the situation awareness program. Please refer to Section 5 for the detailed description of these NERC programs and Fig. 13 for the investments on the five programs in different NERC regions over time.

3. Methodology

In this section, data preprocessing is conducted first. Next, the candidate predictors and resilience related responses are defined. Lastly, the feature selection methods used in this work are reviewed.

3.1. Data preprocessing

A blackout (also called a power outage) is generally defined as the loss of electricity to end users in a large area for a considerable duration [20]. Momentary power outages is not considered in this work. Consistent with IEEE Standard 1366-2012 [21], the sustained power outages lasting more than 5 min are considered. The power outages due to severe weather-induced events from 2007 to 2018 are

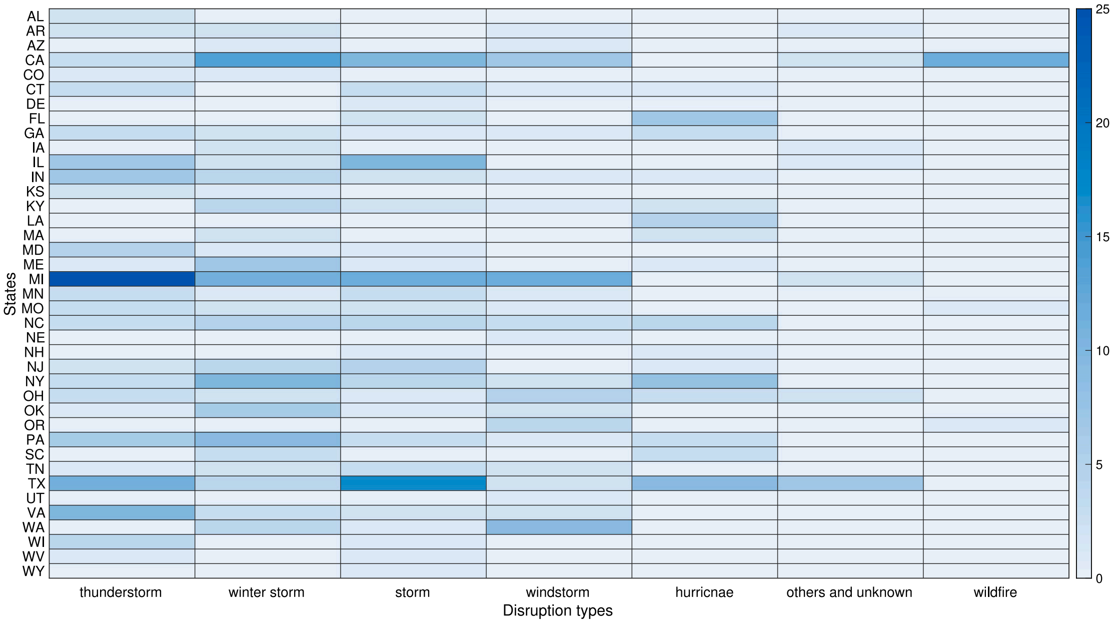

Fig. 2. The frequency of severe weather in each state of the U.S. from 2007 to 2018.

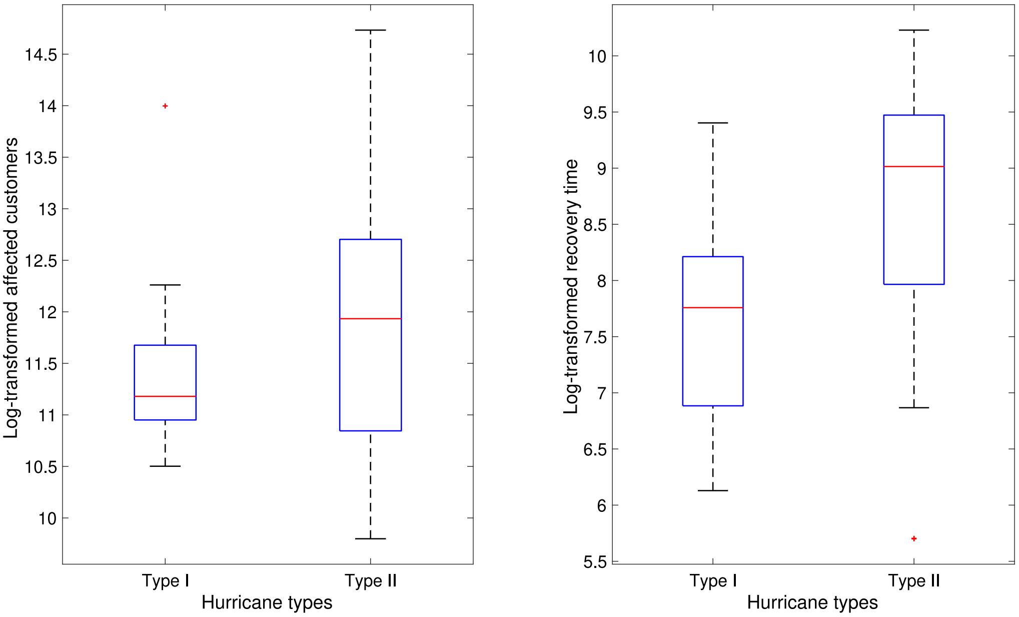

Fig. 3. The boxplot of hurricane types for the number of customers affected and the recovery time from 2007 to 2018 in the U.S. (Left) the number of customers affected is larger when subjected to the type-II-hurricane. (Right) the recovery time is longer when subjected to the type-II-hurricane.

analyzed. In addition, the following data wrangling is done to ensure the data quality. (i) Some observations are identified as outliers and removed. For example, for some observations, the number of customers affected is only 0 or 2 while the recovery time is large. These abnormal observations are probably caused by data entry errors. (ii) Dummy variables are created for the categorical climate regions. (iii) For disruption types, wildfires are incorporated into others and unknown as the sample size is small for the wildfire type. (iv) Dummy variables are created for the disruption types and climate categories, respectively.

Some highly correlated variables in Table 7 are removed to avoid collinearity. Firstly, the redundant variables are removed. For simplicity, the three types indicate residential, commercial and industrial in this paper. (i) For the group of electricity price, the total price in the state is a weighted average of the three types of price. The industrial electricity price is identified as a redundant variable and removed. (ii) For the group of electricity consumption, the total electricity consumption in the state consists of the three types of electricity consumption. The industrial electricity consumption and its percentage are identified as redundant variables and removed. (iii) For the group of the demographic information, the annual total number of customers served in the state consists of the numbers of the three types of customers served. Similarly, the number of industrial customers served and its percentage are removed. (iv) For the group of the geographic information, the percentage of the water area in the state is identified as a redundant variable since the total area in the state consists of the land area and the water area.

Furthermore, the correlation coefficient ρ for each pair of independent variables are examined. A value of the correlation near ±1 indicates that the two variables are highly correlated. Therefore, some variables with high correlation coefficients are removed. (i) For the

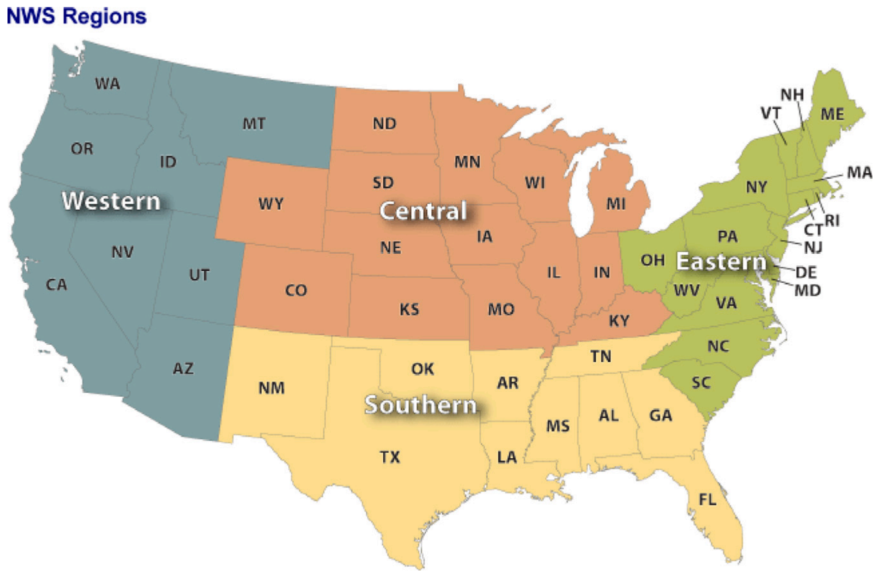

Fig. 4. The national weather service regions in the U.S.

Table 1
Predictors used in our feature selection analysis and their notation.

Dimension	Groups		Predictors	Notation
Extrinsic disruptions	X1	Disruption type	HURRICANE_I	X11
			HURRICANE_II	X12
			WINDSTORM	X13
			THUNDERSTORM	X14
			WINTER STORM	X15
			STORM	X16
	X2	Climate category	COLD	X21
			WARM	X22
			ONI	X23
Intrinsic capabilities	X3	Electricity price	RES.PRICE	X31
			COM.PRICE	X32
	X4	Consumption	RES.SALES	X41
			COM.SALES	X42
			RES.PERCEN	X43
			COM.PERCEN	X44
	X5	Demographic	POPPCT_URBAN	X51
			POPPCT_UC	X52
			POPDEN_URBAN	X53
			POPDEN_UC	X54
			POPDEN_RURAL	X55
			TOTAL.CUSTOMERS	X56
			RES.CUST.PCT	X57
	X6	Geographic	AREAPCT_URBAN	X61
			AREAPCT_UC	X62
			PCT_LAND	X63
			PCT_WATER_INLAND	X64
			REGION_CENTRAL	X65
			REGION_SOUTHERN	X66
			REGION_EASTERN	X67
Effectiveness of recovery	X7	Economics	PC.REALGSP.STATE	X71
			PC.REALGDP.USA	X72
			PC.REALGSP.REL	X73
			PC.REALGSP.CHANGE	X74
			UTIL.CONTRI	X75
	X8	Technical	STANDARDS	X81
			COMPLIANCE	X82
			ASSESSMENTS	X83
			TRAINING	X84
			AWARENESS	X85

group of electricity price, the total price variable is removed as it is highly correlated with the residential price and the commercial price, with ρ = 0.95 and 0.97 respectively. (ii) For the group of electricity consumption, the total electricity consumption is removed as it is highly correlated with the residential consumption and the commercial consumption, with $\rho = 0.97$ and 0.96 respectively. (iii) For the group of the demographic information, the annual population variable is removed as it is highly correlated with the annual total number of customers served, with ρ = 0.99. The numbers of the two types of customers served (residential and commercial) are removed as they are both highly correlated with the annual total number of customers served, with $\rho = 0.99$. In addition, the percentage of the commercial customers served is removed as it is highly correlated with the percentage of the residential customers served, with ρ = −0.95. (iv) For the group of economic outputs, the real GSP contributed by utility industry, the real GSP contributed by all industries, and the ratio of the state utility's income to the U.S. utility's income are removed as they are highly correlated with the annual total number of customers served, with ρ = 0.97, 0.97 and 0.95 respectively.

3.2. Candidate predictors

After some preprocessing which deleted some predictors that exhibit collinearity, the final candidate predictors (in groups) in our analysis are shown in Table 1. There are 39 candidate predictors which have been grouped into 8 natural groups, referring to Table 7 for the definitions of all the notations.

3.3. Candidate responses

To analyze resilience, we look into two key measures associated with each blackout and recovery of power systems, i.e., the number of customers affected $nCust.A$ and the time needed for recovery $rTime$. $rTime$ is usually measured in a relevant time unit, such as minutes, hours or days. The unit of recovery time in our work is minutes. Fig. 5 shows the histograms for the two response variables, based on the historic data from 2007 to 2018. The histograms show that the distributions are skewed to the right, i.e. heavy-tailed, so we take the log-transformed $nCust.A$ and $rTime$ as our final response variables.

3.4. Feature selection

Consider the linear regression problem with p predictors. Assume the predictors could be naturally split into J non-overlapping groups,

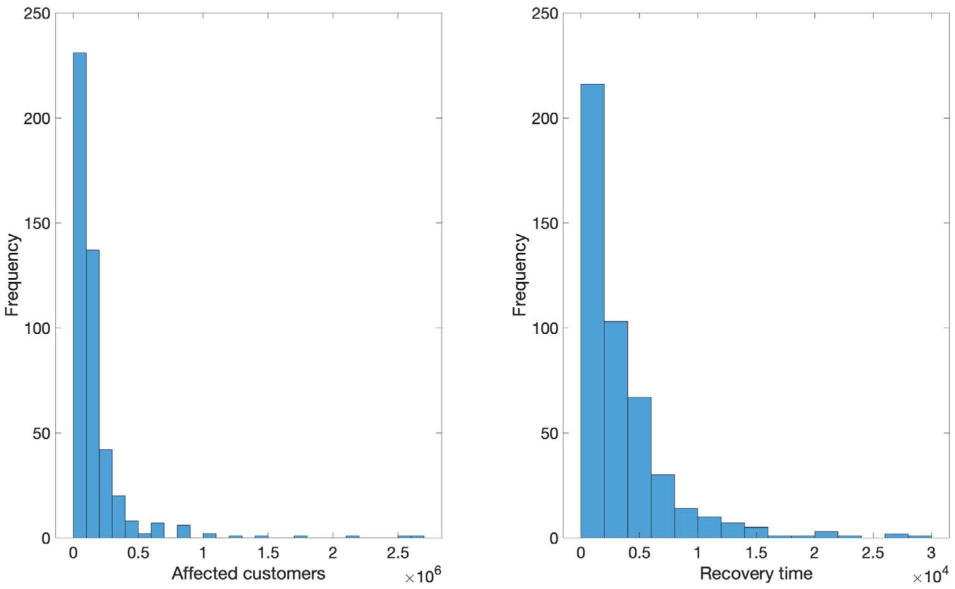

Fig. 5. Histograms for the number of affected customers and recovery time from 2007 to 2018.

and the model is represented as

$$\mathbf{y} = \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{\beta}_j + \varepsilon,$$

where $\mathbf{y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}_j$ is the $n \times d_j$ design matrix of the d_j predictors in the jth group, $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jd_j})' \in \mathbb{R}^{d_j}$ is the $d_j \times 1$ vector of regression coefficients of the jth group, and $\boldsymbol{\varepsilon}$ is an error vector that is uncorrelated with the predictors, with second moment bounded. Denote the elements of $\boldsymbol{X}_j$ by x_{ijk}, which is the value of kth predictor in the jth group for the ith subject. Without loss of generality, the predictors are standardized prior to fitting such that $\sum_i x_{ijk} = 0$ and $n^{-1} \sum_i x_{ijk}^2 = 1$, and they can be transformed back to the original scale once all the models are selected, if needed. The response variables are also centered with zero mean. In this study, we have 39 candidate predictors belonging to 8 natural groups, i.e., $p = 39$ and $J = 8$. The response $\mathbf{y}$ denotes either the number of customers affected or the recovery time. We are interested in selecting both important groups and factors that have effect on the U.S. power system resilience, thus group selection methods and bi-level selection methods will be applied.

3.4.1. Group selection

For group selection methods, the group LASSO (gLASSO) [22], the group SCAD (gSCAD) and the group MCP (gMCP) [17] are used in this study. The gLASSO poses penalty on the weighted l_2-norm of coefficients of a grouped covariates, which selects either the whole group or none of the factors, providing a sound way for our problem. However, the gLASSO does not possess the selection consistency property, and it usually tends to over-fit the model, i.e., selecting more groups than the true model. The gSCAD and gMCP are two l_2-norm concave group selection methods which generalize SCAD [15] and MCP [16] to a group manner. They may perform better than the gLASSO in finding the true model as SCAD and MCP are consistent in variable selection under mild conditions.

(1) gLASSO For a vector $\mathbf{v} \in \mathbb{R}^d$, $d \geq 1$, and a positive definite matrix $\boldsymbol{R}$, we denote $\|\mathbf{v}\|_2 = (\mathbf{v}'\mathbf{v})^{1/2}$ and $\|\mathbf{v}\|_{\boldsymbol{R}} = (\mathbf{v}'\boldsymbol{R}\mathbf{v})^{1/2}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_J')'$, where $\boldsymbol{\beta}_j \in \mathbb{R}^{d_j}$. Given the $d_j \times d_j$ positive definite matrix $\boldsymbol{R}_j$, the gLASSO solution $\hat{\boldsymbol{\beta}}(\lambda)$ is defined as a minimizer of

$$\frac{1}{2n}\left\|\mathbf{y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j\right\|_2^2 + \lambda \sum_{j=1}^{J} c_j \left\|\boldsymbol{\beta}_j\right\|_{\boldsymbol{R}_j}, \tag{1}$$

where $\lambda \geq 0$ is the tuning parameter, c_j is used to adjust for the group sizes and a reasonable choice is $c_j = \sqrt{d_j}$.

The gLASSO penalty can be represented as $\lambda c_j \left\|\boldsymbol{\beta}_j\right\|_{\boldsymbol{R}_j} = \rho(\left\|\boldsymbol{\beta}_j\right\|_{\boldsymbol{R}_j}; c_j\lambda)$, where $\rho(t;\lambda) = \lambda t$ for $t \geq 0$, and we can extend λt to any convex function of t, to result a class of convex penalties.

(2) gSCAD and gMCP A different class of group selection methods is based on the following criterion [17]

$$\frac{1}{2n}\left\|\mathbf{y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j\right\|_2^2 + \sum_{j=1}^{J} \rho(\left\|\boldsymbol{\beta}_j\right\|_{\boldsymbol{R}_j}; c_j\lambda, \gamma), \tag{2}$$

where $\rho(t; c_j\lambda, \gamma)$ is concave in t, and γ is a tuning parameter that may be used to modify ρ. Two example functions for ρ are as follows. The SCAD penalty [15] is defined as

$$\rho(x; \lambda, \gamma) = \lambda \int_0^{|x|} \min\{1, (\gamma - t/\lambda)_+/(\gamma - 1)\} dt, \gamma > 2.$$

The MCP [16] is defined as

$$\rho(x; \lambda, \gamma) = \lambda \int_0^{|x|} (1 - t/(\lambda\gamma))_+ dt, \gamma > 1,$$

where for any $a \in \mathbb{R}$, a_+ denotes its positive part, that is, $a_+ = a1_{\{a \geq 0\}}$. By applying the SCAD and MCP penalties to (2), we obtain the l_2-norm gSCAD and l_2-norm gMCP, respectively.

3.4.2 Bi-level selection

For bi-level selection methods, the composite MCP (cMCP) [23] and the group exponential LASSO (gEL) [24] are used to select both important groups as well as factors within-group. Compared to l_2-norm gMCP, cMCP avoids over-shrinkage by allowing covariates to grow large and allows groups to remain sparse internally. Similar advantages are also enjoyed by gEL over l_2-norm gLASSO.

(1) cMCP Huang et al. [23] proposed a bi-level selection framework, in which grouped penalties are defined as an outer penalty ρ_O applied to a sum of inner penalties ρ_I. The penalty applied to a group of predictors is

$$\rho_O\Big(\sum_{k=1}^{d_j} \rho_I(|\beta_{jk}|)\Big),$$

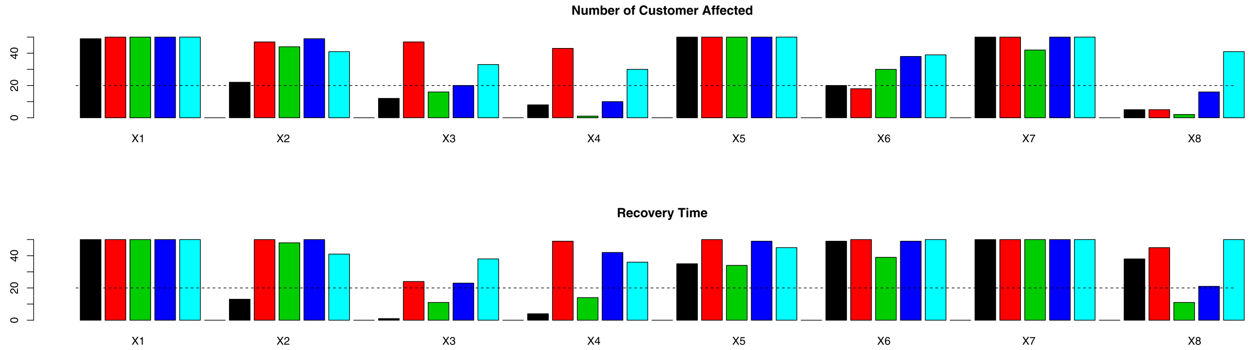

Fig. 6. Group selection results. Y axis denotes the selection frequency, and X axis denotes the 8 groups, including X1 (disruption type), X2 (climate category), X3 (electricity price), X4 (consumption), X5 (demographic), X6 (geographic), X7 (economics) and X8 (technical). For each group, the selection frequencies are for gEL (black), gLASSO (red), gMCP (green), gSCAD (dark blue) and cMCP (light blue). For the bi-level methods (gEL and cMCP), the maximal frequencies within group are presented. The dashed line represents the threshold 20.

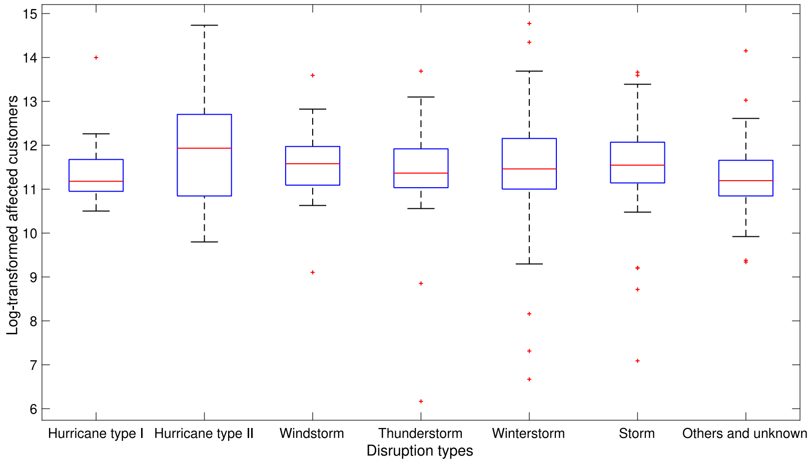

Fig. 7. The box-plot of the log-transformed affected customers and the disruption types.

where β_{jk} is the kth member of the jth group. The cMCP uses the MCP as both the outer and inner penalties

$$\frac{1}{2n}\left\|\mathbf{y}-\sum_{j=1}^{J}\mathbf{X}_j\boldsymbol{\beta}_j\right\|_2^2+\sum_{j=1}^{J}\rho_{\lambda,\gamma_O}\left(\sum_{k=1}^{d_j}\rho_{\lambda,\gamma_I}(|\beta_{jk}|)\right), \tag{3}$$

where ρ is the MCP penalty and the tuning parameter of the outer penalty, γ_O, is chosen to be $d_j\gamma_I\lambda/2$.

(2) gEL Huang et al. [17] proposed a general penalized criterion by applying concave penalties to the l_1 norm of a group

$$\frac{1}{2n}\left\|\mathbf{y}-\sum_{j=1}^{J}\mathbf{X}_j\boldsymbol{\beta}_j\right\|_2^2+\sum_{j=1}^{J}\rho(\left\|\boldsymbol{\beta}_j\right\|_1|\lambda). \tag{4}$$

Breheny [24] proposed the gEL under the criterion, in which the inner penalty $\rho_I(\cdot)$ is the lasso penalty, and the outer penalty is defined as the exponential penalty

$$\rho(x;\lambda,\gamma)=\frac{\lambda^2}{\gamma}\left\{1-\exp\left(-\frac{\gamma x}{\lambda}\right)\right\},$$

where ρ is defined on $[0,\infty)$, and λ and γ are parameters.

3.5 Some comments on the feature selection methods

Theoretically, the methods mentioned above have differences. For example, on one hand, the gLASSO does not enjoy the selection consistency as in gSCAD and gMCP; on the other hand, the loss function in gLASSO is convex in the parameters so that the solution is the global minimizer, while the loss functions in gSCAD and gMCP are not convex so that the solution could be a local minimizer. Furthermore, the numerical performances of all these methods are reasonably well. Therefore, it is really challenging to pick the “best” selection method, and it is hard to compare these methods in the real application. Instead, we include these methods all together for cross validation [25], i.e., if a group is selected by some or all these methods, we can be more confident about the selected group. All the methods mentioned above are implemented in the R package “grpreg”.

4 Application to power systems

Applying the feature selection approaches to the whole data set only results whether a predictor is selected or not, which may be affected by small disturbance. To enhance the robustness of feature selection, we adopt the random subsampling method, in which we randomly pick 70%–80% of the whole sample to perform feature selection, and repeat the procedure for a sufficient number of times, saying 50. Then the importance of the predictors could be ranked by the selection frequency, which is also more informative than selected or not. Similar approaches and settings can be found in [26] and other literature. The algorithm works as follows.

(1) Randomly choose 70% of the data and apply gLASSO, gMCP, gSCAD, gEL and cMCP to select the predictors based on the selected subset.
(2) Repeat step (1) 50 times. Record the selection frequency of all predictors and claim that the predictor is “selected” if the selection frequency is at least 20.

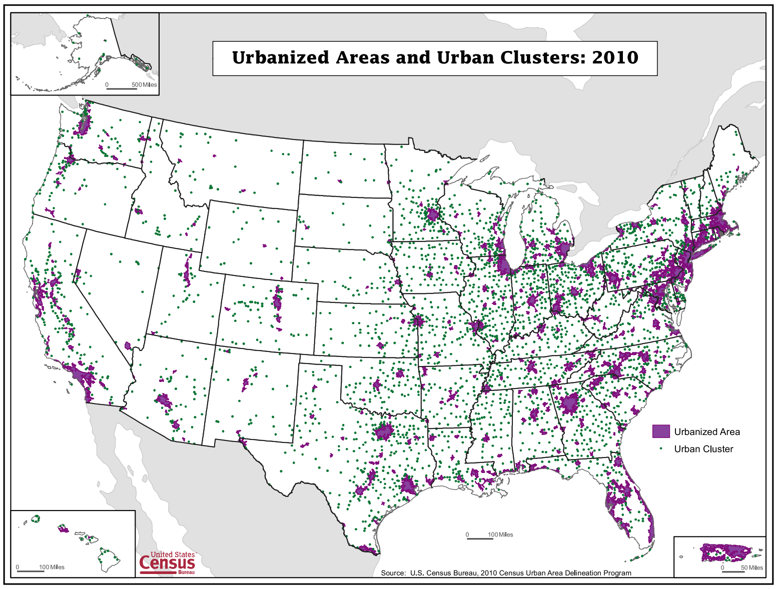
Urbanized Areas and Urban Clusters: 2010
0 500 Miles
0 100 Miles
0 100 Miles
0 50 Miles
Urbanized Area
Urban Cluster
United States Census Bureau
Source: U.S. Census Bureau, 2010 Census Urban Area Delineation Program

Fig. 8. Urbanized Areas and Urban Clusters: 2010.

Table 2
Top five groups identified for the two responses.

nCust.A			*rTime*		
gLASSO	gMCP	gSCAD	gLASSO	gMCP	gSCAD
X1,X5,X7	X1,X5	X1,X5,X7	X1,X2,X5,X6,X7	X1,X7	X1,X7,X2
X2	X2	X2		X2	X5,X6
	X7			X6	
				X5	

Table 3
Predictors selected with top five different frequencies for the two responses.

nCust.A		*rTime*	
gEL	cMCP	gEL	cMCP
X52,X55,X75	X12,X55,X75	X11,X12,X13,X15,X72	X12,X15,X63,X66,X72
X12,X51	X52	X14,X16,X62,X63,X66,X75	X11
X57	X11	X65	X13
X11	X51	X57	X75
X54	X13	X73	X16

The full selection results for *nCust.A* and *rTime* are summarized in Tables 8 and 9 in Appendix A. It is expected that different selection methods yield different results, and we highlight the predictors which are selected at least 20 times by all the methods. We also present the group selection frequencies in Fig. 6.

In the following Sections 4.1 and 4.2, we focus on examining how the top-selected predictors contribute to resilience. Before detailed analysis, we summarize the top five (or less) groups in Table 2 identified by gLASSO, gMCP and gSCAD. The number of presented groups may be less than five because other groups are not "selected", and the groups in the same row are selected with the same frequency. For factors identified by the bi-level selection methods gEL and cMCP, Table 3 presents the predictors with top five different frequencies.

4.1 Selection results for *nCust.A*

We first make a summary on the selection results for *nCust.A* in Table 2: (i) the groups of disruption type (X1) and climate category (X2) from the extrinsic disruption dimension, the demographic group (X5) from the intrinsic capabilities dimension, and the group of economics (X7) from the effectiveness of recovery dimension are selected by gLASSO, gMCP and gSCAD simultaneously, while X1 and X5 are ranked as the most important groups; (ii) within the selected groups, the predictors selected by both gEL and cMCP are highlighted in Table 8, while the top-selected ones are presented in Table 3. Please refer to Tables 1 and 7 for the definitions of all the notation.

Within X1, the type-II-hurricane (X12) is top selected. Fig. 7 presents the box-plots of *nCust.A* with different disruption types, showing that *nCust.A* tends to be larger when subjected to the type-II-hurricane, which is consistent with the discussion in Section 2. As an example, Hurricane Harvey was a devastating Category 4 hurricane that made landfall on Texas (TX) in August 2017, affecting more than 1 million customers.

Within X5, the percentage of urban population (X51), the percentage of urban clusters population (X52) and the population density of rural areas (X55) are ranked as the most important predictors by both gEL and cMCP. Naturally, *nCust.A* during disruptions is closely related to the population factors in the region. The Census Bureau defines urban areas as densely developed territory, encompassing residential, commercial, and other non-residential urban land uses, with at least 2500 people. There are two types of urban areas: urbanized areas and urban clusters. The urbanized area is defined as an area with a population of 50,000 or more, and the urban cluster is defined as an area containing at least 2500 and less than 50,000 people. The map of urbanized areas and urban clusters in 2010 in the U.S. is shown in Fig. 8. X51 and X52 represent the percentage of the total population living in urban areas and urban clusters. The rural area encompasses all population, housing, and territory not included within an urban area, with fewer than 2500 residents. X55 is defined as the rural population divided by the arable land area. The scatter plots of *nCust.A* and X51, X52 and X55 are shown in Fig. 9. The result shows that *nCust.A* tends to be larger in a region with higher X51, and/or in a region with lower X52. Fig. 9 also shows that *nCust.A* tends to be slightly larger in a region with lower X55.

Furthermore, the interaction between these demographic factors and *nCust.A* in each state is given in Fig. 10. It is found that *nCust.A* tends to be larger in the states such as California (CA), New York (NY), and Florida (FL), which have higher percentage of urban population,

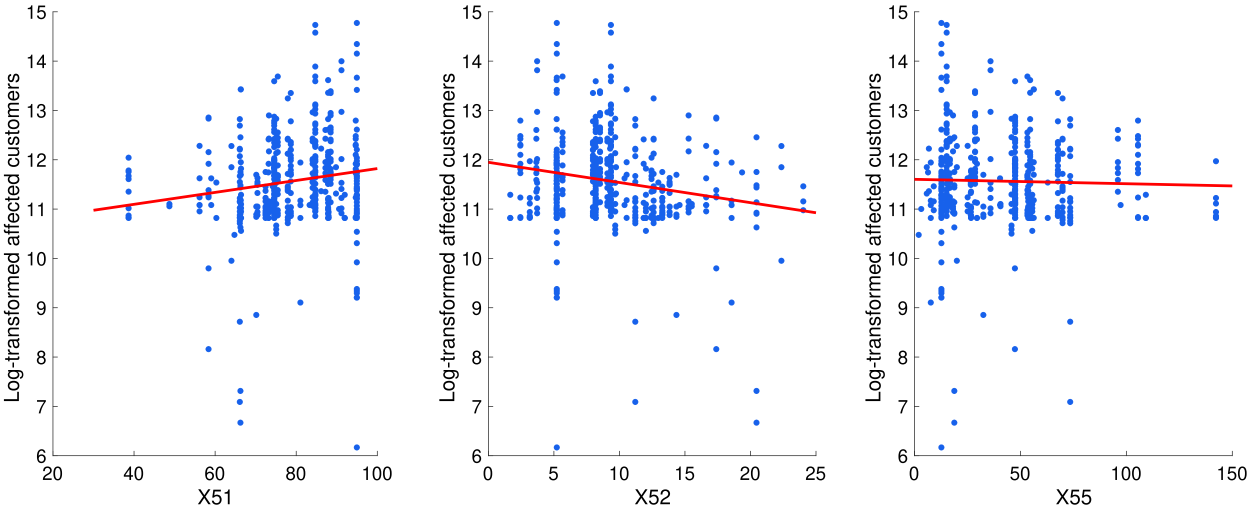

Fig. 9. The scatter plots of the log-transformed $nCust.A$ and the percentage of the urban population in the state (X51), the percentage of the urban clusters population in the state (X52), and the population density of the rural areas (X55).

shown on the left panel in the first row of Fig. 10. Higher urban population often leads to higher proportions of metropolitan areas, higher electricity demand and restricted tree trimming activities, which make power systems in these regions to experience cascading failures more likely. In addition, as discussed before, California (CA) experienced frequent disruptions, and New York (NY) and Florida (FL) experienced frequent hurricanes. To give some power outages examples in which $nCust.A$ is large in these states: (i) the January 2008 North American storm complex caused more than 2 million people to lose power in California (CA); (ii) the 2014 San Diego County wildfires resulted in more than 1 million people lost power; (iii) Hurricane Matthew in 2016 caused more than 1 million people to lose power in Florida (FL). On the other hand, though the percentage of urban population in these states is high, the percentage of urban clusters population in these states is low, which means the percentage of population in urbanized areas is high. Therefore, $nCust.A$ tends to be larger in the states with lower percentage of urban clusters population, shown on the right panel in the first row of Fig. 10. It is also found that $nCust.A$ tends to be slightly larger in the states such as California (CA) and Texas (TX), which have lower population density of the rural areas, shown on the left panel in the second row of Fig. 10.

In a nutshell, the results show that the power system tends to be less resilient in highly urbanized areas with large population, whereas more satellite townships improves resilience against weather-induced disruptions. It suggests that a more distributed population while maintaining a minimum density is a better design in terms of achieving higher resilient power systems for urban planning.

We also look at the normalized performance loss defined as $L = nCust.A/nCust.T$, where $nCust.T$ is the total customers served in the area. The selection results for the performance loss are shown in Table 4, in which the top selected groups are the same as $nCust.A$. The factors X12, X52, X55 are top selected, which are also consistent with the selection results for $nCust.A$. For simplicity, we mainly present the results discussion on $nCust.A$ which represents the direct loss from disruptions.

4.2 Selection results for $rTime$

We first make a summary on the selection results for $rTime$ in Table 2: (i) the groups of disruption type (X1) and climate category (X2) from the extrinsic disruption dimension, the demographic group (X5) and the geographic group (X6) from the intrinsic capabilities dimension, and the group of economics (X7) from the effectiveness of recovery dimension are selected by gLASSO, gMCP and gSCAD simultaneously, while X1 and X7 are ranked as the most important groups (Table 2); (ii) within the selected groups, the predictors selected by both gEL and cMCP are highlighted in Table 9, while the top-selected ones are presented in Table 3.

Table 4
Top five groups and predictors identified for the performance loss.

Group selection			Bi-level selection	
gLASSO	gMCP	gSCAD	gEL	cMCP
X1,X5,X7	X5,X7	X1,X5,X7	X12,X56,X75	X52,X55,X56,X75
X2	X1	X2	X11,X52	X12
	X2		X55	X11
			X14	X14
			X15	X73

Within X1, the type-II-hurricane (X12) and the winter storms (X15) are top selected. The box-plots of $rTime$ with different disruption types are shown in Fig. 11, showing that $rTime$ tends to be longer when subjected to X12 and X15. As discussed before, the type-II-hurricane tends to result in prolonged power outages. Winter storms usually bring freezing rain, sleet, snow, and high winds, damaging power lines and equipment. The $rTime$ tends to be long as the repair may not be timely due to extreme weather, or road maintenance and closures. It may be also due to extensive damage, such as tree falling on power lines. Among the states, California (CA), Michigan (MI) and New York (NY) experienced frequent winter storms, seeing Fig. 2. For example, the January 2010 North American winter storms caused California (CA) to lose power up to 236.5 h.

Within X7, the per capita real GDP in the U.S. (X72) and the utility's contribution to the total GSP in the state (X75) are top selected. X72 is defined as the total economic output of a country divided by the population and adjusted for inflation. On average, we found that the power outage duration tends to be shorter when X72 gets higher, shown on the left panel in Fig. 12. Higher X72 indicates a better economic health, thus the power system tends to receive more investment for hardening and more recovery effort after disruptions, which will improve resilience. On the contrary, the U.S. experienced prolonged power outages in 2009 when X72 is the smallest between year 2007 and 2018 due to global financial crisis.

X75 is defined the as percentage of the total real GSP that is contributed by the utility industry in the state. According to the Bureau of Economic Analysis (BEA), the utilities sector comprises establishments engaged in the provision of the following utility services: electric power, natural gas, steam supply, water supply, and sewage removal.

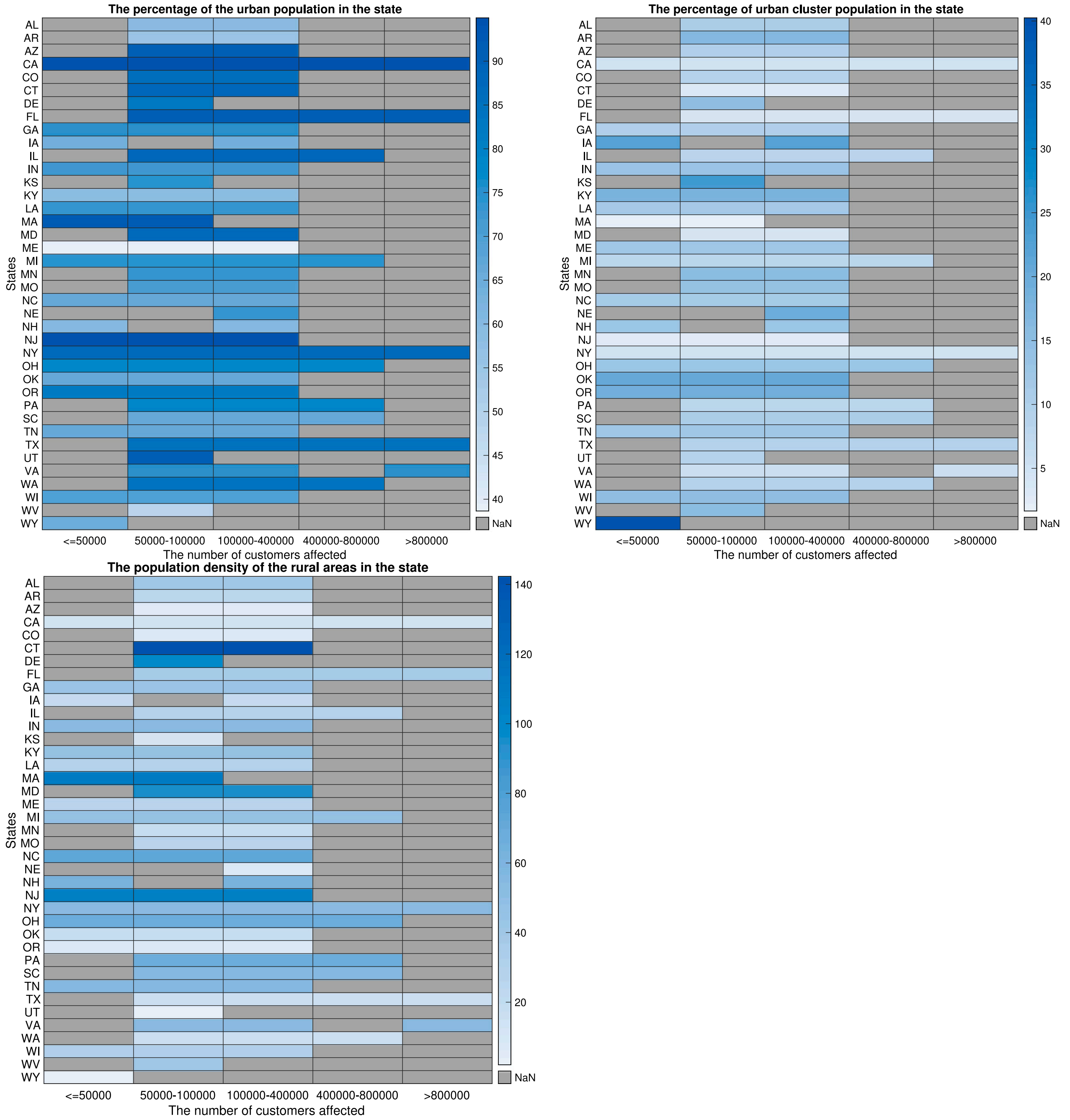

Fig. 10. The interaction between *nCust.A* and the percentage of the urban population in the state (X51), the percentage of the urban clusters population in the state (X52), and the population density of the rural areas (X55).

For power systems, the specific activities associated with the utility services include electricity generation, transmission, and distribution. On average, we found that the power outage duration tends to be longer in a region with higher X75, such as Texas (TX), Michigan (MI), Arkansas (AR) and Louisiana (LA), shown in the middle in Fig. 12. High X75 means that the utility may actively build new infrastructure and construct new capacity, which may affect the investment on regular maintenance and operational activities. Some built-in redundancy capabilities may not be available while constructing new buildings and capacity, which may result in a less resilient system. On the other hand, the post-disaster rebuilding may lead to higher utility output and employment and boost the total GSP growth, which will exhibit a positive association between power outage duration and the utility's contribution. In addition, from the extrinsic disruption dimension, the power systems in Texas (TX) and Louisiana (LA) were exposed to hurricanes impacts frequently which may result in prolonged outages. For example, Hurricane Ike caused Texas (TX) to lose power up to 461 h, and Louisiana (LA) to lose power up to 340 h.

Within X6, the percentage of land area in state (X63) is ranked as one of the most important predictors, which is in line with the result by Mukherjee et al. [13]. X63 is defined as the land area divided by the total area, where the total area in the state includes both land area and water area. Low X63 in a state means that it is more likely to have high percentage of urban population, for example, the states like Massachusetts (MA), Maryland (MD) and Florida (FL). Furthermore, the percentage of population in urbanized areas is high in these states, which indicates that these states tend to have high proportions of metropolitan areas and high commercial activities. As discussed before,

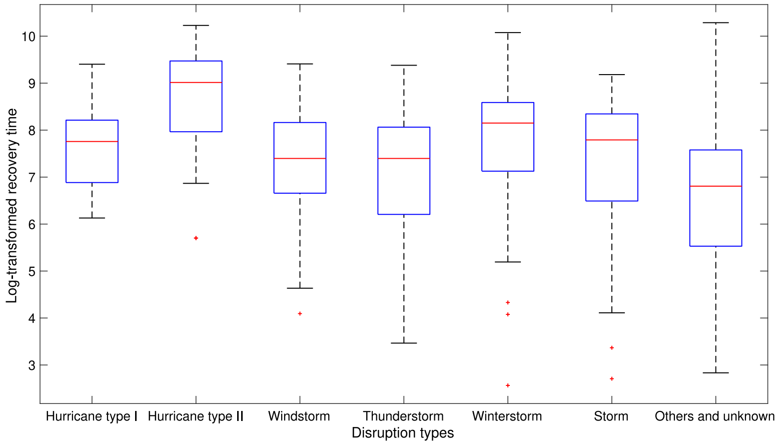

Fig. 11. The box-plot of the log-transformed $rTime$ and the disruption types.

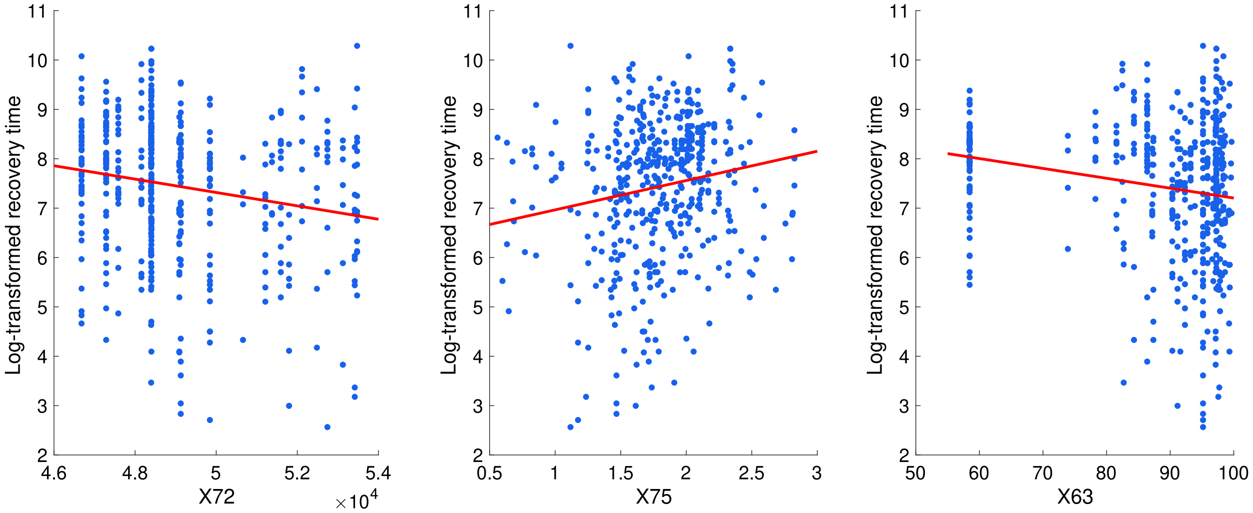

Fig. 12. The scatter plots of the log-transformed $rTime$ and the per capita real GDP in the U.S. (X72), the utility's contribution to the total GSP in the state (X75), and the percentage of land area in the state (X63).

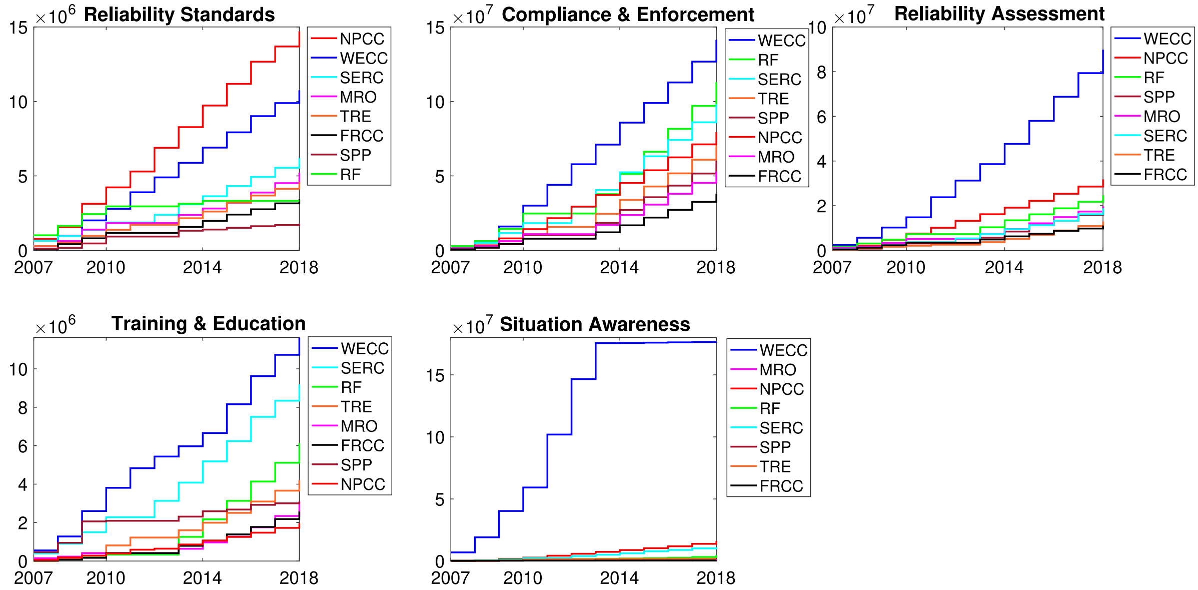

Fig. 13. The cumulative plots of the investments for the five NERC programs in eight NERC regions from 2007 to 2018.

Table 5
Selection frequencies at 0.25, 0.5 and 0.75 quantiles.

	Customers affected			Recovery time		
	0.25	0.5	0.75	0.25	0.5	0.75
X11	2	7	6	13	16	1
X12	11	14	14	38	46	50
X13	36	13	6	24	19	2
X14	7	8	14	45	36	31
X15	13	15	14	46	50	50
X16	44	12	7	15	19	4
X21	20	10	7	33	18	7
X22	6	9	6	12	16	3
X23	49	24	10	45	44	27
X31	9	10	7	22	23	15
X32	44	20	26	31	34	11
X41	11	14	30	21	27	8
X42	3	4	20	9	22	7
X43	23	10	10	34	40	27
X44	39	30	6	39	27	23
X51	45	28	38	40	48	47
X52	50	50	32	37	39	24
X53	7	6	4	14	31	15
X54	25	13	9	35	28	7
X55	50	28	14	24	13	3
X56	38	11	5	38	21	9
X57	34	24	15	38	39	11
X61	37	13	6	33	23	31
X62	15	16	7	41	49	38
X63	19	14	7	50	50	50
X64	35	19	21	38	30	24
X65	9	8	6	17	15	2
X66	4	6	3	37	29	21
X67	11	13	7	14	13	21
X71	19	3	7	12	23	12
X72	32	12	14	50	50	50
X73	5	5	5	15	28	13
X74	21	14	14	33	30	22
X75	50	49	34	50	50	49
X81	25	13	10	38	36	35
X82	27	14	13	49	50	25
X83	42	13	12	32	30	39
X84	33	42	20	27	29	35
X85	45	16	13	34	36	47

the power systems are more likely to experience cascading failures in these states, which may result in prolonged power outages. Among all the states, Michigan (MI) has the lowest percentage of land area (58.5%). Surrounded by water, Michigan (MI) experienced more frequent disruptions compared to other states, such as thunderstorms and windstorms. The high winds and thunderstorms may knock trees and limbs across power lines and break utility poles and equipment frequently, resulting in prolonged power outages. Therefore, on average, we find that the power outage duration tends to be longer in a region with lower X63, shown on the right in Fig. 12. The southern climate region (X66) is also top selected for $rTime$. Most of the southern coastal regions of the U.S. are exposed to hurricane impacts, which may result in prolonged power outage durations, such as Florida (FL), Texas (TX), and Louisiana (LA). In addition, Texas (TX) also experienced frequent thunderstorms and storms. Therefore, the power outage duration tends to be longer in the southern climate region in the U.S.

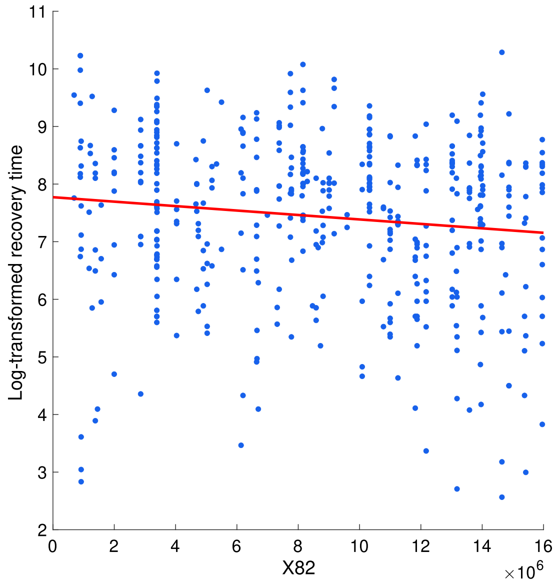

Fig. 14. The scatter plots of the log-transformed $rTime$ and the investment on the compliance and enforcement program (X82).

In a nutshell, for enhancing resilience, our study suggests that utility companies should be well prepared for hurricanes with high scales and winter storms, most of which are predictable a few days in advance. Utility companies need to set a high priority on proactive maintenance activities and ensure redundancy capabilities when expanding economic activities. In addition, utility companies need to conduct tree-trimming programs regularly in areas close to power lines for states with low percentage of land areas which are usually highly urbanized.

5 Variable selection under quantile regression

As shown in Fig. 5, the distributions of the two response variables tend to have heavy tails. Shen and Tang [27] found that the blackout size of the U.S. power system are power-law distributed. It means that minor outages are most frequent, and high impact disturbances are rare, thus the predictors could be substantially different. Therefore, we consider feature selection in quantile regression at different levels in this section. We allow different variables to be selected at different quantiles, so we can capture the difference.

We use the LASSO method for quantile regression at τ = 0.25, 0.5, 0.75, where τ = 0.25 may indicate minor outages, τ = 0.5 indicates median outages, and τ = 0.75 indicates major outages. For quantile regression with LASSO penalty, we use the function “cv.rq.pen” in the R package “rqPen”, selecting the λ by cross-validation. The feature selection results under quantile regression for the three response variables are summarized in Table 5, where the numbers denote the selection frequencies among 50-time random splits, while the top-selected predictors are presented in Table 6.

For 0.25- and 0.5-quantiles of $nCust.A$, the percentage of urban clusters population (X52) is the top selected predictor, which indicates

Table 6
Top-selected predictors at 0.25, 0.5 and 0.75 quantiles.

Customers affected			Recovery time		
0.25	0.5	0.75	0.25	0.5	0.75
X52,X55,X75	X52	X51	X63,X72,X75	X15,X63,X72,X75,X82	X12,X15,X63,X72
X23	X75	X75	X82	X62	X75
X51,X85	X84	X52	X15	X51	X51,X85
X16,X32	X44	X41	X14,X23	X12	X83
X83	X51,X55	X32	X62	X23	X62

Table 7
Full predictors.

Groups	Predictors	Description
Disruption type	HURRICANE_I	Hurricane Type I
	HURRICANE_II	Hurricane type II
	WIND	Windstorms
	THUNDERSTORM	Thunderstorms
	WINTER STORM	Winter storms/ice storms/snow storms
	STORM	Other storms
Climate category	COLD	Cold episode of the climate
	WARM	Warm episode of the climate
	ONI	The Oceanic Niño Index (ONI)
Electricity price	RES.PRICE	Residential electricity price (cents/kilowatt-hour)
	COM.PRICE	Commercial electricity price (cents/kilowatt-hour)
	IND.PRICE	Industrial electricity price (cents/kilowatt-hour)
	TOTAL.PRICE	Average electricity price in the state (cents/kilowatt-hour)
Electricity consumption	RES.SALES	Residential electricity consumption (megawatt-hour)
	COM.SALES	Commercial electricity consumption (megawatt-hour)
	IND.SALES	Industrial electricity consumption (megawatt-hour)
	TOTAL.SALES	Total electricity consumption in the state (megawatt-hour)
	RES.PERCEN	Percentage of residential electricity consumption (%)
	COM.PERCEN	Percentage of commercial electricity consumption (%)
	IND.PERCEN	Percentage of industrial electricity consumption (%)
Demographic	POPULATION	Population in the state in a year
	POPPCT_URBAN	Percentage of the urban population (%)
	POPPCT_UC	Percentage of the urban clusters population (%)
	POPDEN_URBAN	Population density of the urban areas (persons per square mile)
	POPDEN_UC	Population density of the urban clusters (persons per square mile)
	POPDEN_RURAL	Population density of the rural areas (persons per square mile)
	RES.CUSTOMERS	Annual number of residential customers served
	COM.CUSTOMERS	Annual number of commercial customers served
	IND.CUSTOMERS	Annual number of industrial customers served
	TOTAL.CUSTOMERS	Annual number of total customers served
	RES.CUST.PCT	Percent of residential customers served (%)
	COM.CUST.PCT	Percent of commercial customers served (%)
	IND.CUST.PCT	Percent of industrial customers served (%)
Geographic	AREAPCT_URBAN	Percentage of the land area of the urban (%)
	AREAPCT_UC	Percentage of the land area of the urban clusters (%)
	PCT_LAND	Percentage of land area in state (%)
	PCT_WATER_TOT	Percentage of water area in state (%)
	PCT_WATER_INLAND	Percentage of inland water area in state (%)
	REGION_CENTRAL	Central climate region in US
	REGION_SOUTHERN	Southern climate region in US
	REGION_EASTERN	Eastern climate region in US
Economics	PC.REALGSP.STATE	Per capita real GSP in the state (2009 chained U.S. $)
	PC.REALGDP.USA	Per capita real GDP in the US (2009 chained U.S. $)
	PC.REALGSP.REL	Per capita state real GSP/per capita US real GDP
	PC.REALGSP.CHANGE	Percentage change of per capita real GSP (%)
	UTIL.REALGSP	Real GSP contributed by utility industry (2009 chained U.S. $)
	TOTAL.REALGSP	Real GSP contributed by all industries (2009 chained U.S. $)
	UTIL.CONTRI	Utility's contribution to the total GSP in the state (%)
	PI.UTIL.OFUSA	State utility's income/US utility's income (%)
Technical	STANDARDS	Investment on reliability standards program (U.S. $)
	COMPLIANCE	Investment on compliance & enforcement program (U.S. $)
	ASSESSMENT	Investment on reliability assessment and performance analysis (U.S. $)
	TRAINING	Investment on training and education program (U.S. $)
	AWARENESS	Investment on situation awareness program (U.S. $)

that the percentage of urban clusters population affects minor and medium power outages, with a small or medium $nCust.A$. The population density of rural areas (X55) is also top selected for 0.25-quantile. For 0.75-quantile of $nCust.A$, the percentage of urban population (X51) is the top selected predictor, which affects the major power outages with a large $nCust.A$. As discussed in Section 4.1, X51, X52 and X55 are top selected ones for $nCust.A$ under mean regression, thus the selection results under quantile regression tally with those under mean regression.

For $rTime$, the percentage of land area in state (X63), the per capita real GDP (X72) and the utility's contribution to the total GSP in the state (X75) are top selected for all the three quantiles. Besides, the type-II-hurricane (X12) is top selected for 0.75-quantile, which indicates that the type-II-hurricane has a great impact on the long power outages. As discussed in Section 4.2, X72, X75, X63 and X12 are selected as the top predictors for $rTime$ under mean regression. So the selection results under quantile regression tally with those under mean regression.

Furthermore, the investment on the Compliance and Enforcement program (X82) is top selected at both 0.25- and 0.5-quantiles of $rTime$, which indicates that this investment has significant impacts on short and medium power outages. The Compliance and Enforcement program belongs to the NERC investments, which are conducted to improve the reliability and security of the bulk power system. The NERC investments include five programs: (i) the Reliability Standards program (X81), aiming to develop quality reliability standards in a timely manner that are effective, clear, consistent and technically sound; (ii) the Compliance and Enforcement program (X82), with the main goal to improve the reliability of the bulk power system by fairly and

Table 8
Results of selection frequencies of the predictors for $nCust.A$.

	gEL	gLASSO	gMCP	gSCAD	cMCP
X11	40	50	50	50	47
X12	49	50	50	50	50
X13	33	50	50	50	44
X14	35	50	50	50	24
X15	33	50	50	50	41
X16	22	50	50	50	25
X21	17	47	44	49	41
X22	17	47	44	49	24
X23	22	47	44	49	40
X31	0	47	16	20	0
X32	12	47	16	20	33
X41	0	43	1	10	4
X42	0	43	1	10	1
X43	8	43	1	10	30
X44	0	43	1	10	5
X51	49	50	50	50	45
X52	50	50	50	50	49
X53	34	50	50	50	10
X54	39	50	50	50	23
X55	50	50	50	50	50
X56	37	50	50	50	12
X57	43	50	50	50	26
X61	15	18	30	38	2
X62	15	18	30	38	16
X63	14	18	30	38	25
X64	15	18	30	38	27
X65	15	18	30	38	39
X66	15	18	30	38	10
X67	20	18	30	38	33
X71	1	50	42	50	6
X72	17	50	42	50	20
X73	0	50	42	50	5
X74	4	50	42	50	4
X75	50	50	42	50	50
X81	5	5	2	16	41
X82	0	5	2	16	8
X83	0	5	2	16	13
X84	1	5	2	16	19
X85	3	5	2	16	19

Table 9
Results of selection frequencies of the predictors for $rTime$.

	gEL	gLASSO	gMCP	gSCAD	cMCP
X11	50	50	50	50	49
X12	50	50	50	50	50
X13	50	50	50	50	48
X14	49	50	50	50	21
X15	50	50	50	50	50
X16	49	50	50	50	45
X21	1	50	48	50	23
X22	1	50	48	50	15
X23	13	50	48	50	41
X31	0	24	11	23	3
X32	1	24	11	23	38
X41	1	49	14	42	1
X42	1	49	14	42	1
X43	4	49	14	42	36
X44	4	49	14	42	15
X51	11	50	34	49	30
X52	28	50	34	49	28
X53	7	50	34	49	15
X54	3	50	34	49	2
X55	3	50	34	49	4
X56	3	50	34	49	10
X57	35	50	34	49	45
X61	40	50	39	49	11
X62	49	50	39	49	42
X63	49	50	39	49	50
X64	31	50	39	49	4
X65	41	50	39	49	40
X66	49	50	39	49	50
X67	38	50	39	49	6
X71	37	50	50	50	7
X72	50	50	50	50	50
X73	30	50	50	50	31
X74	35	50	50	50	12
X75	49	50	50	50	46
X81	38	45	11	21	50
X82	2	45	11	21	27
X83	4	45	11	21	11
X84	5	45	11	21	15
X85	2	45	11	21	17

consistently enforcing compliance with NERC standards; (iii) the Reliability Assessment and Performance Analysis program (X83), aiming at assessing, measuring and investigating historical trends and future projections to ensure bulk power system reliability; (iv) the Training and Education program (X84), which ensures personnel operating the bulk power system are well-trained and certified to operate the system reliably; and (v) the Situation Awareness program (X85), which aims to spread awareness on the dependencies and facilitate communication among subsystems to promote collaboration in face of disruptions [1]. NERC includes eight regions: Florida Reliability Coordinating Council (FRCC), Midwest Reliability Organization (MRO), Northeast Power Coordinating Council (NPCC), ReliabilityFirst (RF), SERC Reliability Corporation (SERC), Southwest Power Pool (SPP), Texas Reliability Entity (TRE) and Western Electricity Coordinating Council (WECC). The cumulative investments for the five programs in the eight NERC regions from 2007 to 2018 are shown in Fig. 13, which clearly shows that a large portion of the financial support goes to X82. Compared to other programs, some regions have received increasing financial support in this program. As discussed in [1], the increasing investment on this program is expected to improve the resilience. For example, among various NERC regions, the resilience in the NPCC region has become better based on the historical data from 2012 to 2016. Therefore, from an overall perspective, we conclude that the power system tends to be more resilient in a region with higher and increasing investment on the Compliance and Enforcement program. In addition, Fig. 14 shows that $rTime$ tends to be shorter in a region with higher investment on X82.

6 Conclusion

Based on the historical data from 2007 to 2018, we applied several feature selection methods to identify and rank the key predictors affecting power system resilience in the U.S., under both mean regression and quantile regression. (i) The disruption types from the extrinsic disruptions dimension have significant impacts on the resilience of power systems, especially the type-II-hurricane. Utility companies are suggested to be well prepared for hurricanes with high scales. (ii) For the number of customers affected, the demographic group from the intrinsic capabilities dimension is top selected, where the percentage of urban population, the percentage of urban clusters population and the population density of rural areas are the top selected factors under both the mean and quantile regressions. From an overall perspective, the power system tends to be less resilient in more urbanized areas, whereas more resilient in a region with more satellite townships. (iii) For the recovery time, the group of economics from the effectiveness of recovery dimension is top selected, where the per capita real GDP of the U.S. and the utility's contribution to the total GSP are the top selected components under both the mean and quantile regressions. The power system tends to be less resilient with smaller per capita real GDP, and/or with larger utility's contribution to the total GSP. Utility companies are suggested to set a high priority on proactive maintenance activities and ensure redundancy capabilities when expanding economic activities for resilience enhancement. The geographic group from the intrinsic capabilities dimension is also selected, where the percentage of land areas in state is the top selected factor under both the mean and quantile regressions. The power system tends to be less resilient in a state with lower percentage of land areas. Utility companies are suggested to conduct maintenance activities and tree-trimming programs regularly in these states for resilience enhancement. (iv) The investment on the Compliance and Enforcement program by NERC from the technical group is top selected under quantile regression, which indicates that the power system tends to be more resilient with larger and increasing investment on this program. Overall, our results provided fundamental insights on understanding power systems resilience, which are important for developing strategies to enhance resilience. To achieve greater resilience, a roadmap and associated tools for designing resilience into power systems will be investigated in future research.

CRediT authorship contribution statement

Lijuan Shen: Conceptualization, Methodology, Formal analysis, Writing - original draft, Visualization, Writing - review & editing. **Yanlin Tang:** Methodology, Software. **Loon Ching Tang:** Conceptualization, Supervision.

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.

Acknowledgments

This work is an outcome of the Future Resilient Systems project at the Singapore-ETH Centre (SEC) and the Intra-CREATE Seed Collaboration Grant (NRF2021-ITS007-0013) supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The research was also partially supported by National Natural Science Foundation of China grant 11871376, Shanghai Pujiang Program 18PJ1409800, Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education.

Appendix A. Full predictors

See Table 7.

Appendix B. Results of selection for $nCust.A$

See Table 8.

Appendix C. Results of selection for $rTime$

See Table 9.

References

[1] Shen L, Cassottana B, Tang L. Statistical trend tests for resilience of power systems. Reliab Eng Syst Saf 2018;177:138–47.

[2] Bruneau M, Chang SE, Eguchi RT, Lee GC, O'Rourke TD, Reinhorn AM, Shinozuka M, Tierney K, Wallace WA, von Winterfeldt D. A framework to quantitatively assess and enhance the seismic resilience of communities. Earthq Spectra 2003;19(4):733–52.

[3] Vugrin ED, Warren DE, Ehlen MA, Camphouse RC. A framework for assessing the resilience of infrastructure and economic systems. In: Sustainable and resilient critical infrastructure systems. Springer; 2010, p. 77–116.

[4] Shen L, Cassottana B, Heinimann HR, Tang LC. Large-scale systems resilience: A survey and unifying framework. Qual Reliab Eng Int 2020;36(4):1386–401.

[5] Henry D, Ramirez-Marquez JE. Generic metrics and quantitative approaches for system resilience as a function of time. Reliab Eng Syst Saf 2012;99:114–22.

[6] Ouyang M, Dueñas-Osorio L. Multi-dimensional hurricane resilience assessment of electric power systems. Struct Saf 2014;48:15–24.

[7] Shafieezadeh A, Burden LI. Scenario-based resilience assessment framework for critical infrastructure systems: Case study for seismic resilience of seaports. Reliab Eng Syst Saf 2014;132:207–19.

[8] Hosseini S, Barker K, Ramirez-Marquez JE. A review of definitions and measures of system resilience. Reliab Eng Syst Saf 2016;145:47–61.

[9] Figueroa-Candia M, Felder FA, Coit DW. Resiliency-based optimization of restoration policies for electric power distribution systems. Electr Power Syst Res 2018;161:188–98.

[10] Liu H, Davidson RA, Apanasovich TV. Statistical forecasting of electric power restoration times in hurricanes and ice storms. IEEE Trans Power Syst 2007;22(4):2270–9.

[11] Nateghi R, Guikema SD, Quiring SM. Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes. Risk Anal Int J 2011;31(12):1897–906.

[12] Wang Y, Chen C, Wang J, Baldick R. Research on resilience of power systems under natural disasters—A review. IEEE Trans Power Syst 2015;31(2):1604–13.

[13] Mukherjee S, Nateghi R, Hastak M. A multi-hazard approach to assess severe weather-induced major power outage risks in the US. Reliab Eng Syst Saf 2018;175:283–305.

[14] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 1996;58(1):267–88.

[15] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc 2001;96(456):1348–60.

[16] Zhang C. Nearly unbiased variable selection under minimax concave penalty. Ann Statist 2010;38(2):894–942.

[17] Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. Statist Sci Rev J Inst Math Statist 2012;27(4).

[18] Mukherjee S, Nateghi R, Hastak M. Data on major power outage events in the continental U.S.. Data Brief 2018;19:2079.

[19] National Oceanic and Atmospheric Administration. Regional headquarters. 2019, https://www.weather.gov/organization/regional (Accessed August 20, 2019).

[20] Gou B, Zheng H, Wu W, Yu X. Probability distribution of power system blackouts. In: 2007 IEEE power engineering society general meeting. IEEE; 2007, p. 1–8.

[21] IEEE-1366. IEEE guide for electric power distribution reliability indices. IEEE Standards Association; 2012.

[22] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol 2006;68(1):49–67.

[23] Huang J, Ma S, Xie H, Zhang C. A group bridge approach for variable selection. Biometrika 2009;96(2):339–55.

[24] Breheny P. The group exponential lasso for bi-level variable selection. Biometrics 2015;71(3):731–40.

[25] Fan Y, Tang Y, Zhu Z. Variable selection in censored quantile regression with high dimensional data. Sci China Math 2018;61(4):641–58.

[26] Tang Y, Wang Y, Wang HJ, Pan Q. Conditional marginal test for high dimensional quantile regression. Statist Sinica 2021+. (in press).

[27] Shen L, Tang LC. Enhancing resilience analysis of power systems using robust estimation. Reliab Eng Syst Saf 2019;186:134–42.