Predictive Model for Power Outages

Vivian Sultan
California State University
College of Business and Economics
Los Angeles, CA 90032
Email: vsultan3@calstatela.edu

***Abstract*—This study aims to address the question "How to predict power outages?" A statistical model in Statistical Package for Social Sciences (SPSS) is used to predict power-outage-event duration and customer calls using a stepwise regression algorithm. The model presented in this study can help advance smart-grid reliability by predicting power outages and taking the necessary steps to prevent them. Future work may involve enhancing the model's success and adding significant predictive variables.**

Keywords—data analytics; power-failure; smart-grid.

I. INTRODUCTION

In a short time, electrical power has become a necessity of modern life. Our work, healthcare, leisure, economy, and livelihood depend on the constant supply of electrical power. Even a temporary power outage can lead to relative chaos, financial setbacks, and possible loss of life. U.S. cities dangle on electricity and, without a constant supply from the power grid, pandemonium would ensue. Power outages can be especially tragic when they endanger life-support systems in hospitals and nursing homes or systems in synchronization facilities such as airports, train stations, and traffic control. The economic cost of power interruptions to U.S. electricity consumers is $79 billion annually in damages and lost economic activity [1]. In 2017, Lawrence Berkeley National Laboratory provided an update, estimating power-interruption costs have increased more than 68% per year since their 2004 study [2].

Many reasons underlie current power failures. Among these reasons are severe weather, damage to electric transmission lines, shortage of circuits, and the aging of the power-grid infrastructure. Severe weather is the leading cause of power outages in the United States [3]. In 2019, weather events as a whole cost U.S. utilities $306 billion: the highest figure ever recorded by the federal government [4].

The aging of the grid infrastructure is another noteworthy reason for power failures. In 2008, the American Society of Civil Engineers gave the U.S. power-grid infrastructure an unsatisfactory grade [5]. They stated in a report that the power-transmission system in the United States required immediate attention. Furthermore, the report mentioned that the U.S. electric-power grid is similar to those of third-world countries. According to the Electric Power Research Institute (EPRI), equipment such as transformers controlling power transmission need to be replaced, as they have exceeded their expected lifespan considering the materials' original design [6].

Electrical outages have three main causes: (1) hardware and technical failures, (2) environment-related, and (3) human error [7]. Hardware and technical failures are due to equipment overload, short circuits, brownouts, and blackouts, to name a few [8]–[10]. These failures are often attributed to unmet peak usage, outdated equipment, and malfunctioning back-up power systems. Environment-related causes for power outages comprise weather, wildlife, and trees that come into contact with power lines. Lightning, high winds, and ice are common causes of weather-related power interruptions. Also, squirrels, snakes, and birds that come in contact with equipment such as transformers and fuses can cause equipment to momentarily fail or shut down completely [8]. As for the third main cause for electrical outages, human error, the Uptime Institute estimated that human error causes roughly 70% of the problems that plague data centers. Hacking can be included in the human-error category [11].

Analytics have been a popular topic in research and practice, particularly in the energy field. The use of analytics can help advance smart-grid reliability through, for example, elucidating a root cause of power failure, defining a solution for a blackout through data, or implementing a solution with continuous monitoring and management. In this research paper, the aim is to unveil the novel use of data analytics in predicting power-outage-event duration and customer calls. As the objective in this research is to advance smart-grid reliability, this paper explores ways to create a predictive model for power outages.

II. DATA SELECTION AND METHODOLOGY

EPRI's data repository includes the primary datasets used to conduct this analysis. The data sets include data from advanced metering systems, supervisory-control and data-acquisition systems, Geographic Information Systems, outage-management systems, distribution-management systems, asset-management systems, work-management systems, customer-information systems, and intelligent electronic-device databases. Access to datasets was provided as part of EPRI's data-mining initiative; the initiative provides a test bed for data exploration and innovation and seeks to solve major challenges faced by the utility industry [12].

To further enhance EPRI's dataset, data from other institutions are collected and aggregated. Specifically, Georgia Spatial Data Infrastructure and the Georgia GIS Clearinghouse are the sources for monthly temperature and precipitation data [13]. Additional data regarding storm events and storm details come from the National Oceanic and Atmospheric Administration website (NOAA) [14]. The data size is about 76,000 outages with 13 attributes.

The first step of the project methodology was to load data files from EPRI's Data Repository along with all aforementioned weather data to ArcGIS, the Geographic Information Systems platform created by ESRI. To streamline and make sure that the process is reusable and repeatable, ModelBuilder tool in ArcGIS is utilized to design data workflow. The workflow models spatially join the 48 map layers of weather data (from the Georgia Spatial Data Infrastructure and the Georgia GIS Clearinghouse website) with the outage map layer provided by EPRI. This serves as a final dataset for the study. Next, data exploration through correlational analysis in SPSS and GeoDa software has been conducted. The final step was to run a stepwise regression in SPSS.

Prior to all statistical analyses, data preparation follows these steps:

- Several variables need expert interpretation. For instance, the following variables: forestry expected pruning staff hours, average standard tree-pruning miles with bucket, average mechanical tree-pruning miles, average climbing tree-pruning miles, and actual pruning staff hours/circuit mile, had missing data. Aligned with the expert's instruction, missing data for these variables is replaced with a zero (0).

- The variable pole age had missing data. Per the instruction of the client's expert, transformer age replaces the missing data on pole age, with the reasoning that poles and transformers are routinely installed in tandem.

- The following variables: average climbing tree pruning miles, average standard tree-pruning miles with bucket, average mechanical tree-pruning miles, and forestry expected pruning staff hours either perfectly correlated (r = 1.00) with each other or nearly perfectly correlated ($r > 0.90$) with each other. Thus, these variables injected multicollinearity issues into the stepwise-regression equations. Due to the high level of multicollinearity, all of these variables are removed from the regression equation except one, forestry expected pruning staff hours.

- A stepwise regression algorithm is employed to create two regression equations. As Vogt [16] notes, researchers use a stepwise regression algorithm to find the "best" equation possible when regressing a dependent variable onto multiple independent variables. In other words, only sstatistically significant predictors of the dependent variable in a stepwise regression.

III. Analysis and Results

A. Descriptive Statistics in SPSS

Figures 1 and 2 were the outcomes of the initial data exploration of power-outage events. Percentages and frequencies were calculated for two main categories of outages: either based on storm event or due to forestry management. The breakdown is shown in Table I.

Ritchey [15] notes that for categorical variables, percentages and frequencies are the appropriate descriptive statistics to report. A statistical summary was calculated for all continuous variables in the sample. These data appear in Table II.

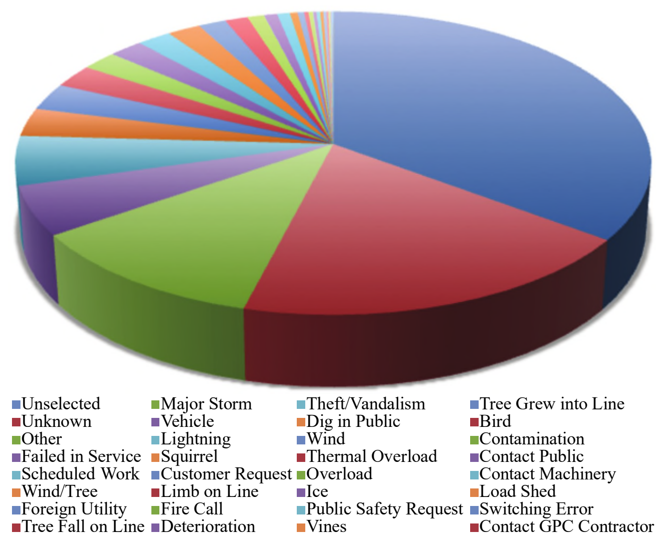

Figure 1. Reported outage events percent count by cause.

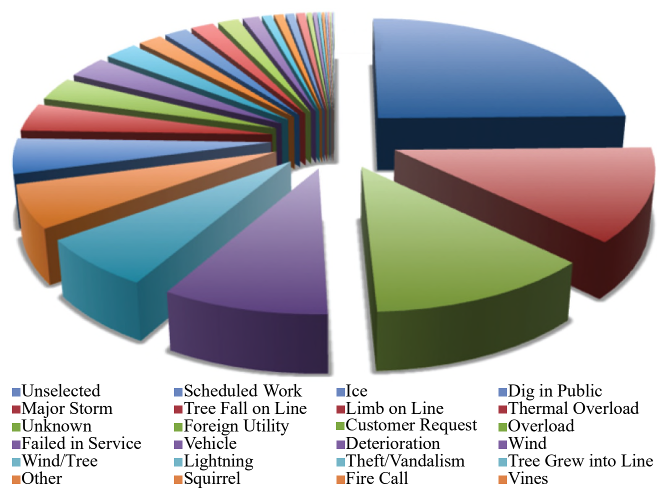

Figure. 2. Reported outage duration by cause.

TABLE I. PERCENTAGES AND FREQUENCIES, STUDY VARIABLES

	Frequency	Percent
Storm event		
Yes	59,078	76.9%
No	17,769	23.1%
Forestry management		
Yes	47,175	61.4%
No	29,672	38.6%
n	76,847	100.0%

Ritchey [15] notes that for continuous variables, means and standard deviations are the appropriate descriptive statistics to report.

B. Correlational Results

Results of correlation analysis in SPSS indicated a strong positive correlation between variables, such as storm events and outage-event duration . As expected, a storm event, precipitation, older poles, higher forestry expected pruning human hours, and higher levels of transformer age increased outage-event duration. A negative correlation between variables appeared between variables such as forestry management and outage-event customer calls. Surprisingly, engaging in forestry management, having older poles, and having lower levels of actual pruning human hours/circuit mile decreases the number of outage-event customer calls.

TABLE II. MEANS AND STANDARD DEVIATIONS, STUDY VARIABLES

Variable	M	SD	Min	Max
Outage event duration (minutes)	89.15	204.81	0	3589
Outage event customer calls	11.18	85.07	0	4888
Temperature (mean)	62.52	13.72	40.25	80.75
Precipitation	4.29	0.66	2.8	5.80
Forestry expected pruning man hours	858.01	882.24	0	3300
Average standard tree pruning miles with bucket	6.63	6.48	0	20.49
Average mechanical tree pruning miles	3.00	2.94	0	9.28
Average climbing tree pruning miles	0.75	0.73	0	2.32
Actual pruning man hours / circuit mile	42.18	36.80	0	157
Transformer age	4.50	1.86	3	8
Pole age	23.90	16.76	3	93

Note: n = 76,847.

Statistically significant correlations were flagged in the correlation table (Table III) in the following manner:

- A single star (*) denotes a significant correlation at the p = .05 alpha level.
- A double star (**) denotes a significant correlation at the p = .01 alpha level.
- No stars means the correlation is not statistically significant, and no relationship exists among the two variables in question.
- An inverse correlation was denoted by a negative sign (-). An inverse correlation means that as one variable increases in value, the other variable decreases in value.
- A positive correlation was denoted by the absence of a negative sign (-). A positive correlation means that as one variable increases in value, the other variable increases in value.

TABLE III. CORRELATION RESULTS

	1	2	3	4	5	6	7	8	9	10	11	12
2	0.09**	1.00										
3	0.08**	0.01	1.00									
4	-0.13**	0.01	0.14**	1.00								
5	0.08**	0.00	-0.06**	-0.37**	1.00							
6	0.01	-0.02**	-0.01	0.00	-0.03**	1.00						
7	0.01**	0.00	0.00	0.00	-0.02**	0.77**	1.00					
8	0.01**	0.00	0.00	0.00	-0.02**	0.81**	0.96**	1.00				
9	0.01**	0.00	0.00	0.00	-0.02**	0.81**	0.96**	1.00	1.00			
10	0.01**	0.00	0.00	0.00	-0.02**	0.81**	0.96**	1.00**	1.00	1.00		
11	0.01*	-0.01**	-0.01	-0.01**	-0.02**	0.91**	0.85**	0.79**	0.79**	0.79**	1.00	
12	0.01*	0.01**	0.01**	0.00	0.02**	-0.13**	-0.11**	-0.12**	-0.12**	-0.12**	-0.11**	1.00
13	0.02**	-0.01**	0.02**	-0.01**	0.03**	0.02**	0.04**	0.03**	0.03**	0.03**	0.03**	-0.03**

Note: 1. Outage-event duration; 2. Outage-event customer calls; 3. Storm event (1 = *yes*); 4. Temperature (mean); 5. Precipitation; 6. Forestry management (1 = *yes*); 7. Forestry expected pruning human hours; 8. Average standard tree-pruning miles with bucket; 9. Average mechanical tree-pruning miles; 10. Average climbing-tree-pruning miles; 11. Actual pruning human hours/circuit mile; 12. Transformer age; 13. Pole age; $^{*}p$ < .05; $^{**}p$ < .01, two-tailed tests.

C. Multiple Linear Regression Results

As Ritchey [15] notes, a multiple linear regression technique is appropriate when the dependent variable is continuous in nature and two or more independent variables are in use. The current circumstances satisfy these criteria. The idea of stepwise regression is to add all independent variables into a regression equation that relates to the dependent variables. Then, the process involves iteratively peruse the regression, removes the variables that are not statistically contribute to the dependent variable. In this paper, we have 2 dependent variables of interest: outage-event customer calls and outage-event duration. These dependent variables will result in two regression equations that will be described below.

Table IV presents the results of the stepwise multiple linear regression of outage-event customer calls onto the several independent predictors. The Omnibus F-test, shown in Table IV, is statistically significant (F = 18.217; df = 5, 76841; p < .001). Thus, the decomposition of effects in the regression model can proceed.

TABLE IV. MULTIPLE LINEAR REGRESSION OF OUTAGE EVENT CUSTOMER CALLS ONTO THE PREDICTORS

Variable	B	$SE(B)$	p
Constant	12.214	1.044	0.000
Forestry management	-4.084	1.515	0.007
Forestry expected pruning man hours	0.004	0.001	0.000
Pole age	-0.072	0.018	0.000
Transformer age	0.499	0.167	0.003
Actual pruning staff hours/circuit mile	-0.068	0.024	0.004
N	76847		
F	18.217		0.000
Adjusted R^2	0.001		

Based on the significance of the table, five variables have been retained using the stepwise regression algorithm. Three of these variables lower the number of outage-event customer calls. Specifically, engaging in forestry management (B = -4.084, p = .007), having older polls (B = -0.072, p < .001), and having lower levels of actual pruning staff hours/circuit mile (B = -0.068, p = .004) decrease the number of outage-event customer calls. Two of the variables raise the number of outage-event customer calls. Specifically, having higher levels of forestry expected pruning staff hours ($B = 0.004$, $p < .001$) and higher levels of transformer age ($B = 0.499$, $p = .003$) increase the number of outage-event customer calls. The adjusted R^2 value was identical to the R^2 value.

Table V presents the results of the stepwise multiple linear regression of outage-event duration onto the several independent predictors. The Omnibus F-Test in Table V is statistically significant ($F = 218.672$; $df = 5, 76841$; $p < .001$). Thus, decomposition of effects in the regression model can proceed.

TABLE V. MULTIPLE LINEAR REGRESSION OF OUTAGE EVENT DURATION ONTO THE PREDICTORS

Variable	*B*	*SE(B)*	*p*
Constant	-62.703	5.511	0.000
Storm event	42.413	1.744	0.000
Precipitation	24.796	1.112	0.000
Pole age	0.243	0.044	0.000
Forestry expected pruning man hours	0.004	0.001	0.000
Transformer age	0.898	0.398	0.024
N	76847		
F	218.672		0.000
Adjusted R^2	0.014		

A count of five variables were retained by the stepwise regression algorithm. All five variables raise the outage-event duration. Specifically, having a storm event ($B = 42.413$, $p < .001$), having precipitation ($B = 24.796$, $p < .001$), having older poles ($B = 0.243$, $p < .001$), having higher forestry expected pruning staff hours ($B = 0.004$, $p < .001$), and higher levels of transformer age ($B = 0.898$, $p = .024$) increase the outage-event duration.

IV. CONCLUSION

This study aimed to address "How to predict power outages." To address the research, A predictive novel model was developed in SPSS to predict the power-outage-event duration and customer calls. A stepwise regression algorithm was used for the two regression equations. The SPSS model presented in this study can help advance smart-grid reliability by predicting power outages and taking the necessary steps to prevent them. Future work may involve enhancing the model's success and adding significant predictive variables. Data analytics can be a major resource of assistance for managing power-failure events.

One limitation of this research is that pole-age data was used as a proxy for infrastructure age and the rest of the equipment data. Future work may involve connecting to virtually any type of streaming data feed and transforming data-analytics applications into frontline decision applications, predicting and updating power-outage incidents as they occur.

From this research, it was concluded that SPSS and GIS tools offers a solution to analyze the electric-grid distribution system. This model provides evidence that SPSS can perform the analysis to predict power-outage-event duration and customer calls. If additional funds and data are made available, one can expand on this analysis to create a custom solution for the utility industry to control and forecast power outages. Data analytics can be a major resource of assistance for electronic-inspection systems, to lower the duration of customer outages, to improve crew-response time, and to reduce labor and overtime costs.

REFERENCES

[1] K. LaCommare and J. Eto, "Understanding the cost of power interruptions to U.S. electricity consumers," https://emp.lbl.gov /sites/all/files/lbnl-55718.pdf

[2] J. Eto, "The national cost of power interruptions to electricity consumers —Revised update," http://grouper.ieee.org/groups/td /dist/sd/doc/2017-01-10 National Cost of Power Interruptions to Electricity Customers - Eto.pdf

[3] President's Council of Economic Advisers and the U.S. Department of Energy's Office of Electricity Delivery and Energy Reliability, "Economic benefits of increasing electric grid resilience to weather outages," http://energy.gov/sites/prod/files/2013/08/f2/Grid Resiliency Report_FINAL.pdf

[4] J. Porter, "The $306 billion question: How to make outage management better?" June 12, 2018, https://www.elp.com/Electric-Light-Power-Newsletter/articles/2018/06/the-306-billion-question -how-to-make-outage-management-better.html

[5] American Society of Civil Engineers, "Infrastructure fact sheet," 2009 http://www.infrastructurereportcard.org/2009/sites/default/files /RC2009_rail.pdf

[6] D. Stone, "It's the electric grid, stupid," September 9, 2011, http://www.thedailybeast.com/articles/2011/09/09/major-power-outage -shows-weakness-of-aging-electric-infrastructure.html

[7] K. Chayanam, "Analysis of telecommunications power outages—Inline," https://etd.ohiolink.edu/rws_etd/document/get/ohiou1125024491 /inline

[8] Westar Energy, "What causes power outages? – working to improve service reliability," https://www.westarenergy.com/outage-causes

[9] Rocky Mountain Power, "Key causes of power outages." https://www .rockymountainpower.net/ed/po/or/kcopo.html

[10] Diesel Service and Supply, "Causes of power failures & power outages," http://www.dieselserviceandsupply.com/Causes_of_Power _Failures.aspx

[11] R. Miller, "How to prevent downtime due to human error," https:// www.datacenterknowledge.com/archives/2010/08/13/how-to-prevent -downtime-due-to-human-error

[12] Electric Power Research Institute, "EPRI distribution modernization demonstration (DMD) data mining initiative," http://smartgrid.epri .com/DMD-DMI.aspx

[13] Georgia, G. I. S. Data Clearinghouse,"About the Clearinghouse." Online at www. gis. state. ga. us/Clearinghouse/clearinghouse. html.

[14] National Oceanic and Atmospheric Administration, Storm events database.", 2018, Online at https://www.ncdc.noaa.gov/stormevents/

[15] F. Ritchey, The statistical imagination: Elementary statistics for the social sciences, 2nd edition. Boston, MA: McGraw-Hill, 2008.

[16] W. P. D Vogt, Dictionary of statistics & methodology: A nontechnical guide for the social sciences, 2nd edition. Thousand Oaks, CA: Sage, 1999.