21, rue d'Artois, F-75008 PARIS
http://www.cigre.org

CIGRE US National Committee
2019 Grid of the Future Symposium

Big Data Framework for Predictive Risk Assessment of Weather Impacts on Electric Power Systems

M. KEZUNOVIC, T. DOKIC
Texas A&M University
USA

SUMMARY

Loss of electric power leads to major economic, social, and environmental impacts. It is estimated that the Annual economic impacts from weather-related electric grid outages in the U.S. result in as high as $150 billion. Due to the high level of environmental exposure of the electric utility overhead infrastructure, the most dominant cause of electricity outages is weather impact. More than 70% of electric power outages are caused by weather, either directly (e.g., lightning strikes to the equipment, trees coming in contact with lines under high wind speeds), or indirectly due to weather-caused increases in equipment deterioration rates or overloading (e.g. insulation deterioration, line overloading due to high temperature causing high demand). This paper illustrate how the impact of severe weather can be significantly reduced, and in some cases even eliminated, by accurate prediction of where faults may occur and what equipment may be vulnerable. With this predicted assessment of network vulnerabilities and expected exposure, adequate mitigation approaches can be deployed.

To solve the problem, variety of approaches have been deployed but none seem to be addressing the problem comprehensively. We are introducing a predictive approach that uses Big Data analytics based on machine learning using variety of utility measurements and data not coming from utility infrastructure, such as weather, lightning, vegetation, and geographical data, which also comes in great volumes, is necessary. The goal of this paper is to provide a comprehensive description of the use of Big Data to assess weather impacts on utility assets. In the study reported in this paper a unified data framework that enables collection and spatiotemporal correlation of variety of data sets is developed. Different prediction algorithms based on linear and logistic regression are used. The spatial and temporal dependencies between components and events in the smart grid are leveraged for the high accuracy of the prediction algorithms, and its capability to deal with missing and bad data. The study approach is tested on following applications related to weather impacts on electric networks: 1) Outage prediction in Transmission, 2) Transmission Line Insulation Coordination, 3) Distribution Vegetation Management, 4) Distribution Transformer Outage Prediction, and 5) Solar Generation Forecast. The algorithms shows high accuracy of prediction for all applications of interest.

KEYWORDS

Asset management, Big Data, insulation coordination, outage management, prediction model, smart grid, solar generation forecast, transformer health, vegetation management, weather impact.

1. INTRODUCTION

The electric utility applications have relied heavily on physical model-based solutions in the past. Such methods are not able to estimate or predict the dynamically changing impacts over time which makes it difficult to assess the unfolding deterioration of power grid infrastructure, anticipate fault location, and predict operating conditions [1]. The advancements in smart grid measurement technologies in recent years and wide availability of measurements coming from multiple domains have enabled the necessary conditions for development of new data-driven solutions.

In recent years, a variety of power system data analytics studies in the literature have incorporated data-driven approaches based on various data mining techniques: regression models [2], clustering and classification [3], support vector machines [4], neural networks [5], deep learning [6], etc. Clustering and classification methods have found their place in event classification applications based on PMU data. Support Vector Machines have proven to be powerful in dynamic stability analysis based on synchrophasor data. Neural network solutions have been used in various applications, e.g., optimal maintenance scheduling and optimal placement of various components in the network. Deep learning techniques are finding their way into various applications for real-time load forecasting and emergency management. Regression models have shown great performance by utilizing historical measurements to predict future events in the network through ether logistic or linear regression.

This paper introduces a survey of Big Data (BD) applications based on the regression models capable of providing outage predictions and illustrates the possible mitigation strategies.

2. USE OF BIG DATA FOR POWER SYSTEM STUDIES

It is important to look into advancements in data analytics and identify the ways they can help improve the reliability of the system with advanced prediction methods. These prediction methods can mitigate outages, improve the resilience of the system, and reduce restoration time and cost:

- With more information coming from the new measurements, and other data collected in many domains surrounding network-related events, the accuracy of algorithms used for power system applications can be improved. For example, the use of weather conditions correlated with lightning, soil and vegetation data can improve the predictive capabilities significantly.
- The predictive capabilities can be used to move the decision-making practice from mostly reactive, which is dominant today, to more proactive one. If we are able to predict outages, we can significantly improve the overall reliability of the system by developing preventive mitigation measures.
- Electric networks have experienced a number of changes in recent years, including the addition of renewable sources such as concentrated and distributed solar and wind generation, and electric vehicle integration. BD Analytics platform can provide an automatic platform to support the dynamics of these changes in the network caused by the variability and uncertainty.

3. PREDICTIVE FRAMEWORK

We Use several applications to illustrate how the proposed BD predictive framework overcomes obstacle such as [7]:

- Data Management: The use of BD introduces challenges in data ingestion, cleansing, curation, and wrangling. Also, storing large volumes of data, dealing with data integration at different temporal and spatial scales, understanding that the data sets may contain bad and missing data, and facing varying uncertainty levels from one set to another creates additional concerns.
- Data Analytics: With the use of the graph-based machine learning methods one can achieve: 1) high accuracy of prediction based on modeling the spatial and temporal interdependencies between variables, Prediction at multiple temporal and spatial scales used in operation and maintenance, 3) robustness to missing and bad data by using variables from the nearby nodes.
- Economic Impact: The BD Analytics enables one to define: 1) User-specific mitigation options for risk reduction, 2) risk minimization using optimization with different objectives and constrains. 3) Evaluation framework for comparison between the legacy solutions and BD solutions

3.1 Transmission Outage Prediction

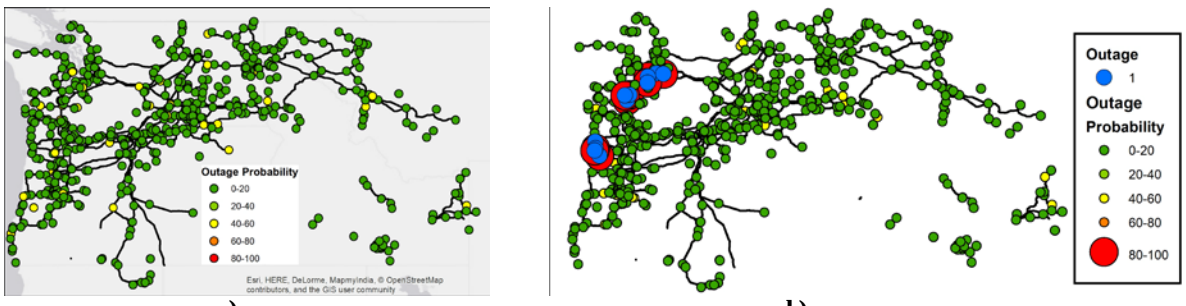
We utilized the knowledge from historical data to provide predictions of weather related outages in the transmission system 1-3 hours ahead. We added the use of spatial embeddings to the input data set [8] to capture the spatial interdependencies between nodes and events. The historical outage data was collected from Bonneville Power Administration (BPA) [9]. The Automated Surface Observing Systems (ASOS) program [10] data was used to collect the historical weather measurements for the following parameters: Wind Direction [degrees], Wind Speed [knots], Wind Gust [knots], Temperature [F], Dew Point [F], Relative Humidity [%], Pressure [mb], Precipitation/Hour [inch], and Present Weather Codes. The National Digital Forecast Database (NDFD) [11] was used to extract the historical weather forecast data that is used for testing of the real-time outage probability mapping system. Three algorithms were implemented and tested: logistic regression (LR), logistic regression with spatial embeddings (LR spatial), and Collaborative Logistic Ensemble Classifier (CLEC) [12]. The results are presented in Table I [8]. We can observe that for all cases spatial solutions have better performances compared to conventional logistic regression algorithm. The CLEC algorithm outperforms both other solutions. Fig. 1 [8] presents the real-time outage probability maps. In ideal case, the predicted probability is high (red color) at the outage locations, and low (dark green color) everywhere else. The following can be observed from the maps in Fig. 1: 1) for the no-outage case, the predicted probability of outages was less than 60%; 2) for the cases with multiple outages in the network, the area with faults had points with high outage probability (over 80%), while the rest of the network had probability lower than 60%.

3.2 Insulation Coordination

We used our method for optimal placement of line surge arresters that minimizes the overall risk of lightning related outages and disturbances, while staying within the required budgetary limits. We model the network and its surrounding impacts using multi-modal weighted graph that uses data coming from various sources. The developed risk model takes into account the accumulated impact of past lightning disturbances in order to produce more accurate estimate of insulator strength, and predicts insulator performances for the future lightning caused overvoltages using Gaussian Conditional Random Fields (GCRF) [13]. Linear programming (LP) [14] is used to find the LSA placement for which the global risk function is minimal. The method has been simulated and tested on section of the network containing 36 substations, 65 transmission lines, with a total of 1590 towers. The historical outage and lightning data for the period of 5 years were observed. The Risk Map is shown in Fig. 2. For each moment in time, it is possible to generate a unique risk map. By averaging the set of risk maps for a period of time it is possible to develop a final risk map on a seasonal or yearly basis. Based on the cumulative risk map produced for a period of one year, and accompanying economic impact, the recommended number of line surge arresters (LSAs) is calculated to be 264, and optimal locations of the LSAs in terms of risk reduction are presented in Fig. 3. The presented configuration of LSAs is expected to reduce overall risk by 72%. This kind of result could help utilities make decision about installation of LSAs in an economically efficient way.

Table I. Prediction performance w.r.t. different evaluation metrics [8].

Model	Acc.	AUC	F1	Bias
LR	0.8467	0.9278	0.8097	0.6821
LR(spatial)	0.8624	0.9292	0.8242	0.7075
CLEC	**0.8919**	**0.9313**	**0.8532**	**0.7685**



a) b)

Figure 1. Probabilities and locations of outages for: a) no outage, b) lightning [8].

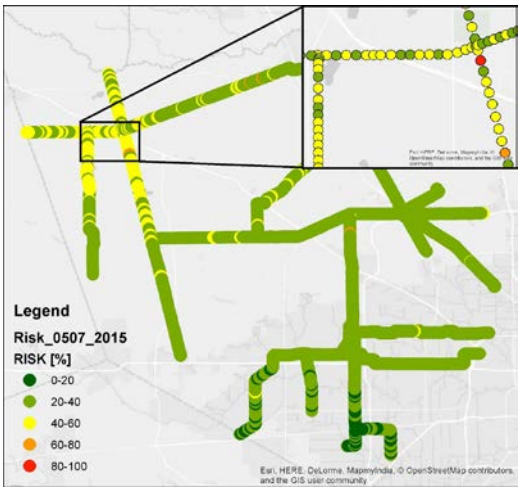


Fig. 2 Risk Map of the Network [14].

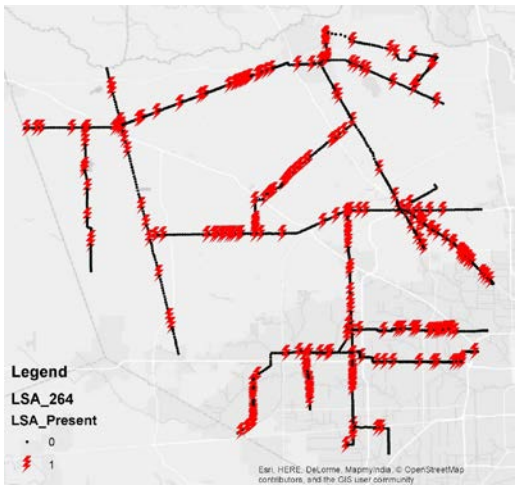


Fig. 3 Locations of 264 LSAs [14].

3.3 Vegetation Management

We introduced the predictive data-driven method for vegetation management in distribution [15]. A model for spatiotemporal correlation of a variety of data is developed, which enables real-time analysis of the vegetation impact on the distribution feeders based on predictive risk maps. Prediction algorithm is based on the GCRF regression predictor. The optimization algorithm is used to find the most cost-effective dynamic tree trimming schedule that minimize the overall risk of the network for each quarter. The benefits of this method are confirmed on an actual utility distribution network in Texas. The area of the analyzed network is ~2,000 km2, containing approximately 200,000 poles, and 120,000 feeders. The historical data was collected from January 2011 up to the end of April of 2016. Over this period, 90% of collected data was used for training of the prediction algorithm, while the remaining 10% of outages at the end of 2015 and beginning of 2016 was used as testing set. The predicted risk map for 02/23/2016 is presented in Fig. 4. We can observe that the predicted risk value on the faulted section was 84%. An example of the developed tree trimming schedule for the first three months in 2016 is presented in Fig. 5. The zones with colors different than black represent the ranges of the network that must be trimmed before the designated deadline.

3.4 Distribution Transformer (DT) Health Assessment

We studied [16] the failures of step down transformers (22.9KV-220V) used in the distribution sectors in South Korea. We collected the data for modeling outage events used for prediction and analysis starting from year 2012 up to year 2018, total of 237 events. The historical outages are extracted for five causes; lighting, tree contact, snow, rain, and dust. Weather parameters considered in this study are: lightning, average temperature, highest temperature, relative humidity, maximum wind speed,

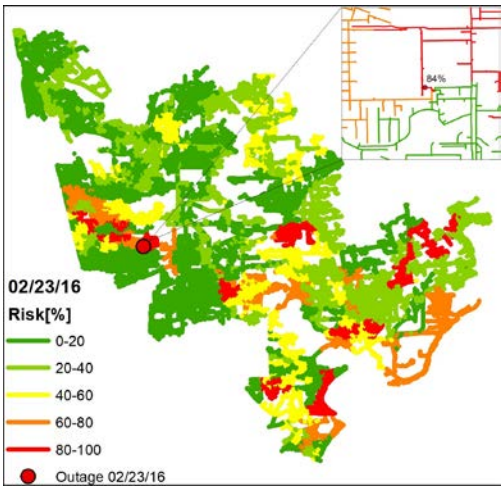


Fig. 4. Risk Map for 02/23/2016 [15].

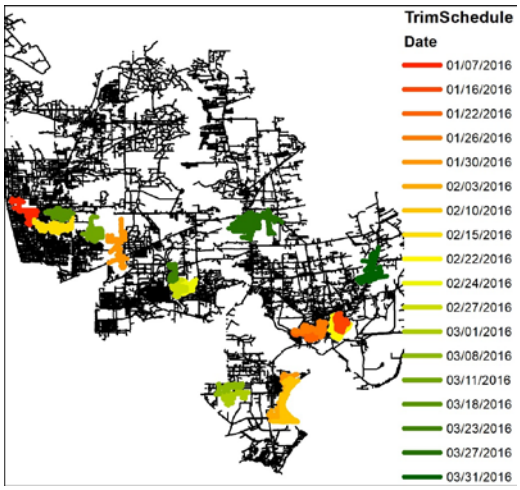
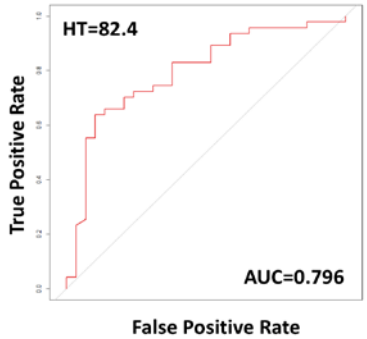
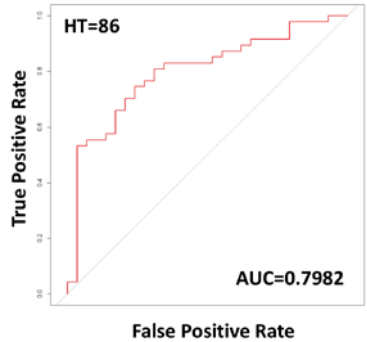


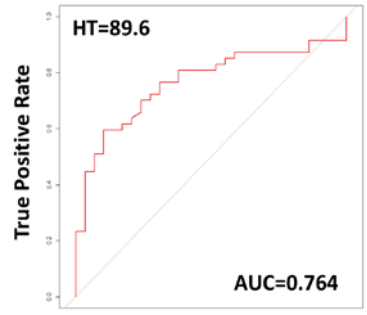
Fig. 5. Quarterly Tree Trimming Schedule [15].



a) HT=82.4℉



b) HT=86℉



c) HT=89.6℉

Figure 6. ROC for classification

wind gust, precipitation. The dates which have outages caused by weather are selected for Y=1 and the dates which don't have any outages are presented as Y=0 and historical weather are extracted. Logistic regression is used for modeling a binary response (i.e., success/fail). This model estimates the probability of the response occurring $P(X) = Pr(Y=1 \mid X)$ through a linear function of explanatory variables X. In this study, it is natural that the response variable Y is a DT failure, i.e., 1 (failure) and 0 (no failure). Specifically, X is $n \times (p+1)$ design matrix where n is the number of observations and p is the number of weather predictors. Naturally, the number of coefficients is eight by seven predictors and an intercept. The corresponding coefficients β of predictors designate the effect of the weather predictors on the probability of DT failure. To evaluate logistic regression, Receiver Operating Characteristics (ROC) graphs and The Area Under Curve (AUC) are used. The historical DT failure data is divided into the testing and training sets. The 90% of the total data set is selected for training. The remaining 10% of the data is used for the testing. The degree of high temperature (HT) is classified into three temperature thresholds such as 82.4°F, 86°F, and 89.6°F in order to make interpretation of HT coefficient precise. The model reported the AUC of 0.796, 0.798 and 0.764 for 82.4°F, 86°F, and 89.6°F, respectively as shown in Fig 6.

4.5 Solar Generation Forecast

As our last example we describe the use of BD analytics prediction for the solar generation based on spatial and temporal embeddings for a Random Forest Regression predictor. The Node2Vec framework was used [17] that learns feature representations for nodes in graphs. To convert the dataset using Node2Vec, a connected graph is created from the solar irradiance grid. In order to achieve this, distances between locations are used as edge weights in a fully connected graph. Then, Node2Vec is used to convert solar data to the new feature space. Two temporal embeddings are created for the observed time period: Hour of the day (0 to 23 value) and Season (winter, spring, summer, and fall). Then, the Random Forest Regression model is applied to predict a value for GHI. Results are presented in Tables 2-4. In the summer model (Table 2), weather measurements from June and July of 2017 are used for training while weather predictions for August 2017 are used for testing. Table 4 shows the results for the summer model. In the winter model (Table 3), weather measurements from October and November of 2017 are used for training, and weather predictions for December 2017 are used for testing. The summer model performs better than the winter model. This is expected due to the higher number of clear sunny days in the summer when the correlation between SZA and GHI is very high. In the global model (Table 4), weather measurements from December and August of 2017 are used for testing and the rest of the months of 2017 are used for training. As expected, this model performs better than the winter model.

Table 2. Results of the summer model

Metric	R^2		MAE	MSE	RMSE
Value	0.91		42.76	8615.9	92.8

Table 3. Results for the winter model

Metric	R^2	MAE	MSE	RMSE
Value	0.85	27.3	5510	71.49

Table 4. Results for the global model

Metric	R^2	MAE	MSE	RMSE
Value	0.89	33.4	7258.9	85.2

5. CONCLUSIONS

Multiple applications related to weather impacts on electricity network have been analyzed: Outage prediction in Transmission, 2) Transmission Line Insulation Coordination, 3) Distribution Vegetation Management, 4) Distribution Transformer Outage Prediction, and 5) Solar Generation Forecast. This survey points out to great potential of predictive methods if the BD is available and properly utilized.

6. ACKNOWLEDGEMENT

Authors would like to acknowledge following individuals for their contributions during the development of presented applications: Prof. Zoran Obradovic, Dr. Ryan Said, Dr. Jelena Gligorijevic, Dr. Djordje Gligorijevic, Martin Pavlovski, Mohammad Alqudah from Temple University in Philadelphia and , and Eun Hui Ko from KEPCO, formerly M.Sc. student at Texas A&M University.

BIBLIOGRAPHY

[1] M. Kezunovic, T. Dokic, "Predictive Asset Management Under Weather Impacts Using Big Data, Spatiotemporal Data Analytics and Risk Based Decision-Making," 10th Bulk Power Systems Dynamics and Control Symposium – IREP'2017, Espinho, Portugal, August 2017.

[2] A. Y. Saber, A. R. Alam, "Short-term load forecasting using multiple linear regression for Big Data" In Computational Intelligence (SSCI), 2017 IEEE Symposium Series on, pp. 1-6.

[3] Wang, Y., et al., "Clustering of electricity consumption behavior dynamics toward Big Data applications," IEEE transactions on smart grid, 7 (5), pp. 2437-2447, 2016.

[4] S. Ye, et al., "Dual-stage feature selection for transient stability assessment based on support vector machine," in Proc. IEEE CSEE, 30, pp. 28–34, 2010.

[5] S. K. Tso, et al., "An ANN-based multilevel classification approach using decomposed input space for transient stability assessment," Elect. Power Syst. Res., 46 (3), pp. 259–266, 1998.

[6] X. Z. Wang, et al., "A multilevel deep learning method for Big Data analysis and emergency management of power system," IEEE Int. Conference on Big Data Analysis, pp.1-5, 2016.

[7] M. Kezunovic, et al., "Predicating Spatiotemporal Impacts of Weather on Power Systems using Big Data Science," Springer Verlag, Data Science and Big Data: An Environment of Computational Intelligence, Pedrycz, Witold, Chen, Shyi-Ming (Eds.), 2017.

[8] T. Dokic, et al., "Spatially Aware Ensemble-Based Learning to Predict Weather-Related Outages in Transmission," HICSS 2019, Maui, Hawaii, January 2019.

[9] Bonneville Power Administration, "Miscellaneous Outage Data and Analysis," [Online] Available: https://transmission.bpa.gov /Business/Operations/Outages/default.aspx

[10] NOAA, "Automated Surface Observing System (ASOS)" [Online] Available: https://www.ncdc.noaa .gov/data-access/land-based-station-data/land-based-datasets/automated-surface-observing-system-asos

[11] National Weather Services, "NDFD," Available: https://www.weather.gov/mdl/ndfd_home

[12] M. Pavlovski, et al. "Generalization-Aware Structured Regression towards Balancing Bias and Variance." IJCAI. 2018.

[13] Radosavljevic, K. Ristovski, Z. Obradovic, "Gaussian Conditional Random Fields for Modeling Patients' Response to Acute Inflammation Treatment," Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013.

[14] M. Kezunovic, et al., "Optimal Placement of Line Surge Arresters Based on Predictive Risk Framework Using Spatiotemporally Correlated Big Data," at CIGRE General Session, Paris, France, Aug. 2018.

[15] T. Dokic, M. Kezunovic, "Predictive Risk Management for Dynamic Tree Trimming Scheduling for Distribution Networks," IEEE Transactions on Smart Grid, Accepted for Publication, in Press, 2018.

[16] E. Hui Ko, et al., "Prediction Model for the Distribution Transformer Failure using Correlation of Weather Data," 5th International Colloquium "Transformer Research and Asset Management" Opatija, Croatia, October, 2019.

[17] A. Grover, J. Leskovec, "node2vec: Scalable Feature Learning for Networks" Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855-864. ACM, 2016.