Paris 2020

Type here your Paper number
6 characters, 9 for joint meetings
(to be centred)

Big Data Analytics for Predictive Lightning Outage Management Using Spatially Aware Logistic Regression Model

M. KEZUNOVIC[1], T. DOKIC[1], Z. OBRADOVIC[2], M. PAVLOVSKI[2], R. SAID[3]
Texas A&M University[1], Temple University[2], Vaisala[3]
USA

SUMMARY

This paper illustrates how Big Data may be used to predict lightning outages in the transmission system. A comprehensive database that contains the necessary information about the historical lightning related events is developed and utilized for training of the prediction algorithm. A Mixture-of-Experts model incorporating multiple spatially aware logistic regression models is utilized to calculate highly accurate short-term predictions 1-3 hours in advance. Lightning strikes to the transmission lines were considered in this study. Different durations of lightning related failures were taken into account, including both temporary and permanent faults. The predictions allow a smart decision-making approach to implementing the proposed mitigation techniques. The model was tested using real utility data. The results demonstrate the capability of the algorithm to predict lightning outage probability with high accuracy for a specific location. The prediction accuracy of the developed algorithm is 0.9370, with the Area Under the Curve being 0.7576. This suggests that the algorithm is good at both predicting high probability for cases with outages, and low probability for cases without outages. The outage probabilities are calculated in real time for every substation and transmission line in the network. Thus, this research provides a significant improvement over the existing studies that aggregate the lightning outage expectancy over a larger geographical area.

KEYWORDS

Big Data, Lightning, Logistic Regression, Outage Prediction, Spatiotemporal Analysis, Weather Forecast, Weather Measurements.

1. INTRODUCTION

In recent years the occurrence of weather-related outages in electric power networks is on the rise. The main factors are changing environmental impacts, and aging infrastructure. Lightning related outages are the second most common weather related outage category, right after the outages caused by a combination of severe winds and vegetation impact.

The Basic Lightning Impulse Insulation Level (BIL) is defined as a voltage at which insulator has 10% probability of a flashover [1]. The BIL value is calculated for the standard atmospheric conditions. The impact of lightning on transmission insulators differs depending on the present atmospheric conditions [2]. Thus, it is important to analyze localized weather parameters to ensure a precise prediction of the lightning impact.

It is not always easy to observe the changes in the insulator lightning performances. Overhead line insulators are exposed to a variety of environmental impacts [3]: lightning strikes, temperature and pressure variations, ultraviolet radiation and ozone, wind impact, rain, humidity, hail, snow, fog, and pollution. In addition, vegetation presence around the line lowers the probability of flashover in the network, a phenomenon called "shielding by trees" [4]. Also, lightning strikes are more likely to affect locations with higher altitude, thus the elevation data is of importance [5]. The tower grounding resistance also has an impact on overvoltage propagation on the line. This resistance is dependent on the type of soil at the tower location. In conclusion, the prediction of probability of lightning outages depends on a variety of factors. To accommodate this challenge, this research collects an extensive dataset including a variety of weather parameters, lightning parameters, and elevation data.

Number of studies are analyzing the risk of the network exposure to lightning outages and available countermeasures [6-9] using different optimization techniques, such as unconstrained nonlinear optimization in [6], multi-objective optimization method based on genetic algorithm in [7], genetic algorithm [8] and linear regression [9] for optimal placement of LSA. For the purpose of estimating probability of a lightning strike, historical lightning data has been used in [10, 11]. All of the mentioned studies use limited amount of data, overlooking the variety of parameters. The studies provide estimation of lightning associated risk on a larger spatial scale by aggregating the impact over large geographical area or averaging it for the component of the same type. This study overpowers the existing solutions by providing precise localized prediction of lightning related risk in the network based on wide variety of data sources discussed next.

2. DATA PRE-PROCESSING

This study uses extensive data sources: 1) Geographical Information System (GIS) data about utility assets (locations of substations, transmission lines, transmission towers, and substation transformers), 2) utility historical outage records, 3) historical weather measurements coming from the land-based weather stations, 4) historical weather forecast data, and 5) elevation data.

Historical data about lightning outages is collected from a utility in the northwest of the USA for a period of 20 years, starting with year 1999, up to the end of year 2018 [12]. A total of 10697 lightning related outages was identified. For each lightning related outage the following parameters were collected from the utility event logs: 1) outage location, 2) outage time and date, 3) operating voltage. The network geographical data is presented in Fig. 1. A total of 639 substations and 686 transmission lines were selected for the study.

The precise locations of lightning strikes and their parameters are collected from the National Lightning Detection Network [13]. Lightning dataset includes following parameters for each lightning strike: 1) location, 2) date and time, and 3) lightning peak current and polarity. The lightning data is spatiotemporally correlated with the historical outage data, where each outage is assigned the lightning event closest to it in time and space. For each location of interest, the number of historical lightning events was extracted in the 10 km radius area around the point.

Historical weather parameters were obtained from the historical land-based weather station data collected by the Automated Surface Observing Systems (ASOS) [14]. The map of locations of 84 weather stations across the network area is presented in Fig. 1. Extracted weather parameters are listed in Table 1. If there was no recorded measurement of a parameter within 1 hour of the targeted time the value was declared missing. Table I lists the fractions of missing data.

Weather forecast data was obtained from the National Digital Forecast Database (NDFD) [15]. The same weather parameters presented in Table 1 were extracted from the NDFD data. In this paper we focus on the short term prediction of lightning outage probability; therefore, the weather forecast for a time interval of 1-3 hours was used. An exception was the precipitation probability which is forecasted every 12 hours. The spatial resolution of forecast data is 5 km.

Elevation data was extracted using Elevation API provided by the Google Maps Platform [16].

3. SPATIOTEMPORAL CORRELATION OF DATA

Prior to feeding the data as inputs to the prediction algorithm, all the data sets have to be spatiotemporally correlated. This includes the extraction of parameters from the ASOS, Outage, and Forecast tables individually as a first step. We create tree datasets, each containing detailed spatial and temporal reference: 1) historical weather data from ASOS, 2) historical outage data from utility, and 3) historical weather forecast from NDFD. The second stage of processing creates training and testing datasets by extracting the measured and forecasted weather parameters for each historical outage. More details about spatiotemporal correlation of data used for this research can be found in [21].

Table I. Missing weather data.

Parameter	Missing %
Temperature [F]	0.15
Dew Point [F]	0.15
Relative Humidity [%]	0.15
Wind Direction [degrees]	0.15
Wind Speed [knots]	0.14
Precipitation [inch/hour]	0.31
Pressure [mb]	0.27
Wind Gust [knots]	0.38
Weather Code	0.34

4. PREDICTION MODEL

4.1. Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{s}_1, t_1, y_1), \ldots, (x_N, \mathbf{s}_N, t_N, y_N)\}$ be a training set of N measurement-outcome pairs $(\mathbf{x}_i, \mathbf{s}_i, t_i, y_i)$ referred to as *labeled* examples. Each example $\mathbf{x}_i \in \mathbb{R}^d$ denotes a measurement vector

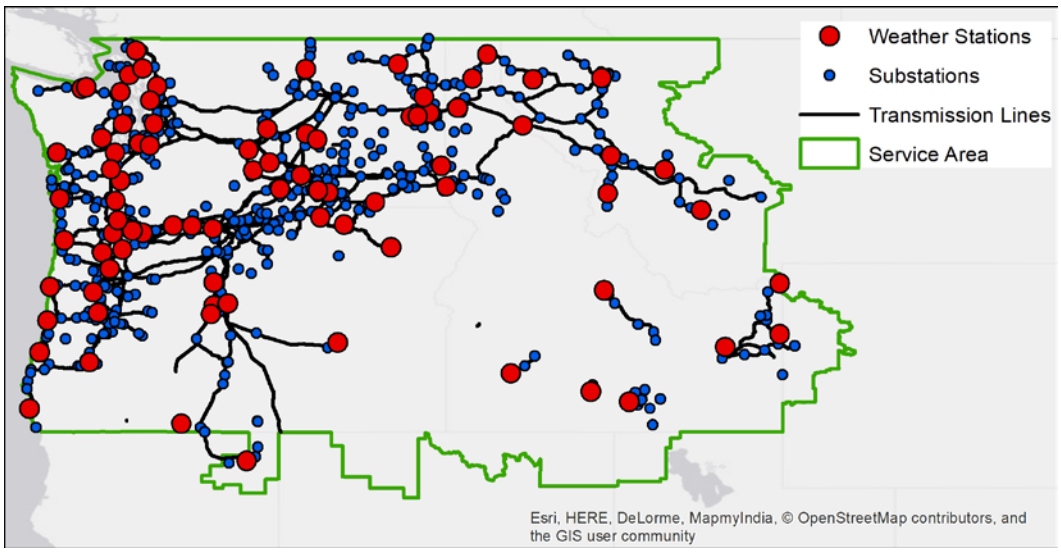


Figure 1. Locations of ASOS weather stations and network components.

containing various weather-related measurements and is associated with 1) a location $\mathbf{s}_i \in \mathbb{R}^3$ (consisted of the latitude, longitude, and elevation) at which the measurements were taken, 2) the measurement time $t_i \in \mathbb{R}$, and 3) a label $y_i \in \{0,1\}$ representing an *event* that indicates whether or not a lightning-induced outage occurred. Note that since multiple measurements were taken at each substation, multiple $\mathbf{x}_i$'s might be associated with the same location coordinates.

The **objective** of this study is to predict the event y (normal operation or a lightning-induced outage), given a vector of weather-related measurement $\mathbf{x}$, location properties $\mathbf{s}$, and measurement time t.

4.2 Spatial Information Embedding

The problem formulated in Section 4.1. boils down to a binary classification task. Therefore, one can simply train a Logistic Regression model to learn the relationship between the events y_i and the substations' measurements $\mathbf{x}_i$. Nevertheless, Logistic Regression being a traditional classification model, assumes the measurements to be independent observations. Namely, it cannot capture the dependencies among the measurements such as the spatial correlations described by the spatial distances between their corresponding substations. Leveraging this information is important since a transmission line leaving a substation is likely to experience an outage in case an outage occurred on a transmission line leaving a nearby substation. To cope with this challenge, a spatial substation graph is constructed and representations of its nodes (substations) are learned based on the spatial distances between them. These representations are referred to as *spatial embeddings*.

Spatial Graph Construction. The spatial graph $G = (V, E)$ is constructed by treating each of the M substations as a separate node v_m $(m = 1, \dots, M)$ in V and initially populating E with edges e_{mp} for each pair of nodes (v_m, v_p). A weight is assigned to each edge e_{mp} as the geospatial distance d_{mp} between the m-th and the p-th substation. The edge weights of the resulting complete graph are binned and the value of the largest increase in bin frequency is detected and used as a threshold for edge removal. In other words, G is sparsified by removing from E the edges whose weights are smaller than a pre-specified threshold.

Node Embedding. Conventional supervised learning models require a set of informative, discriminative, and independent features. Therefore, G needs to be transformed into a set of feature vectors. This can be achieved by learning feature representations, or embeddings, for each node in G. For this purpose, we utilized *node2vec* algorithm [17]. *Node2vec* learns a mapping function $\phi: V \to \mathbb{R}^{d'}$ from nodes to feature representations, where d' is the number of dimensions of the node feature representation. The mapping ϕ is learned such that it maximizes the log-probability of observing a graph neighborhood $N_G(v)$ for a node v conditioned on its feature representation, i.e.

$$\max_{\phi} \sum_{v \in V} \log P\big(N_G(v) \big| \phi(v)\big) = \max_{\phi} \prod_{n \in N_G(v)} P\big(n \big| \phi(v)\big),$$

where a softmax function is used to model the conditional likelihood for every source-neighborhood node pair as:

$$P\big(n \big| \phi(v)\big) = \frac{\exp\big(\phi(n) \cdot \phi(v)\big)}{\sum_{u \in V} \exp\big(\phi(u) \cdot \phi(v)\big)} \ .$$

The partition function in the denominator is expensive to compute, and thus is approximated using negative sampling [18]. The parameters defining ϕ are obtained by optimizing the log-probability above using stochastic gradient ascent and ϕ is used to project each node $v_m \in V$ to a feature

vector $\mathbf{x}_m^s \in \mathbb{R}^{d'}$, for each $m = 1, \dots, M$. Finally, the obtained spatial embedding for the m-th substation is concatenated to the measurement vector $\mathbf{x}_i$ if $\mathbf{s}_i = \mathbf{s}_m$, for all $i = 1, \dots, N$.

Community (Region) Detection. Once a spatial embedding is learned for each node (substation), the original dataset $\mathcal{D}$ is extended to $\mathcal{D}^s = \{([\mathbf{x}_1, \mathbf{x}_1^s], \mathbf{s}_1, t_1, y_1), \dots ([\mathbf{x}_N, \mathbf{x}_N^s], \mathbf{s}_N, t_N, y_N)\}$ in which a spatial embedding $\mathbf{x}_i^s$ is concatenated to each original measurement vector $\mathbf{x}_i$. Note that for simplicity of notation, $\mathbf{x}_i$ will be used to denote the extended feature vectors $[\mathbf{x}_i, \mathbf{x}_i^s], i = 1, \dots, N$. That being done, the M substations are partitioned into K disjoint subsets using the K-Means clustering algorithm. Initially, the cluster centroids $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ are initialized randomly. Thereafter, each measurement vector $\mathbf{x}_i$ is assigned to the cluster for which the distance between its substation and the substation corresponding to the cluster center is minimal, i.e.

$$c_m = \arg\min_{k \in [1,K]} \|\mathbf{s}_m - \boldsymbol{\mu}_k\|_2^2, \quad \text{for each } m = 1, \dots, M;$$

$$\boldsymbol{\mu}_k = \frac{\sum_{m=1}^{M} I(c_m = k)\mathbf{s}_m}{\sum_{m=1}^{M} I(c_m = k)}, \quad \text{for each } k = 1, \dots, K.$$

The two steps above are repeated until convergence. It should be noted that another interpretation of the discovered clusters is to think of them as communities since using the distance between substations' coordinates as a distance metric for the K-Means clustering imposes measurements taken at substations that are nearby (or in the same region) to fall in the same cluster (or community), thus performing an 'implicit', or indirect, community detection.

The entire pipeline, from the spatial graph construction, through the spatial node embedding

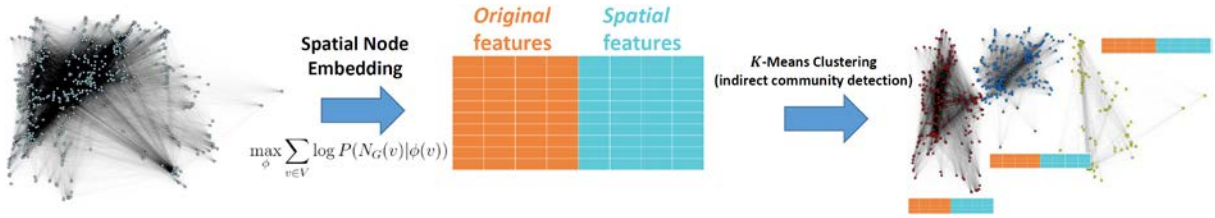


Figure 2. Spatial embedding of the nodes in the spatial graph, followed by an implicit community detection carried out by K-Means clustering of the learned node embeddings.

and up to the subgraph partitioning is illustrated in Figure 2.

4.3 Spatially-Aware Mixture of Logistic Models

Logistic Regression [19], being a probabilistic discriminative classifier, can be directly applied to the task of predicting lightning-induced outages. Given a vector of measurements $\mathbf{x}$, a Logistic Regression model computes the probability of a lightning outage occurrence y in the following manner:

$$P(y_i|\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^{\mathrm{T}}\mathbf{x})},$$

where $\mathbf{w}$ are the model coefficients. Considering this outage probability formulation, a Logistic Regression model is trained by determining the optimal coefficients $\mathbf{w}$ that minimize

$$\mathcal{L} = \sum_{i=1}^{N} y_i \log P(y_i|\mathbf{x}_i; \mathbf{w}) + (1 - y_i) \log\big(1 - P(y_i|\mathbf{x}_i; \mathbf{w})\big).$$

Starting from initial coefficient values (usually zeros), the $\mathcal{L}$ is minimized using gradient-based methods due to its convexity. Upon obtaining the optimal $\mathbf{w}^*$, given an unobserved measurement vector $\mathbf{x}$, its corresponding lightning outage probability is predicted as $P(y = 1|\mathbf{x}; \mathbf{w}^*)$.

Although Logistic Regression can be directly applied to power outage prediction, when applied in a conventional manner, it does not utilize spatial information when predicting outages. On the other hand, a Logistic Regression model trained on spatial embeddings, concatenated to the original features, can utilize the spatial structure of the substations and therefore potentially enhance the model's predictive accuracy. In place of the measurement vectors $\mathbf{x}_i$, one can use the extended feature vectors that contain the spatial embeddings generated using *node2vec* (Section 4.2). This simple extension of the feature space implicitly incorporates **spatial-awareness** in the modeling capacity of Logistic Regression.

A spatial-aware Logistic Regression model may still depend heavily on the overall graphical structure. Slight perturbations in the training data may lead to fluctuations in the model's predictions. In order to **capture relevant spatial substructures** and reduce the variance, and therefore improve the stability of a single Logistic Regression model, multiple Logistic Regression models are employed by generating K disjoint subsets of measurements $\mathcal{D}^1, \ldots, \mathcal{D}^K$ using the partitioning technique from the *Community (Region) Detection* part of Section 4.2. Then, each $\mathcal{D}^k$ is used to train a single Logistic *expert model* $f_{\mathcal{D}^k}$, or simply f_k, described by a separate set of optimal weights $\mathbf{w}^*$. Consequently, the mixture-of-experts architecture incorporates a so-called *gating network* that determines the expert, or blend of experts, whose output is most likely to accurately predict an outage indicator y. The gating outputs are a set of scalar coefficients g_k that weight the contributions of the expert models by minimizing

$$\mathcal{L}_{MEM} = \sum_{i=1}^{N} \log \sum_{k=1}^{K} g_k P(y_i|\mathbf{x}_i; \mathbf{w}_k^*) .$$

Note that the experts' importance coefficients $g_k \geq 0$ and $\sum_k g_k = 1$. Finally, given a measurement vector $\mathbf{x}$, the probability of a lightning outage occurrence based on $\mathbf{x}$ is calculated by taking the weighted combination of the probabilities estimated by the base models, i.e.

$$\mathcal{P}(\mathbf{x}) = \sum_{k=1}^{K} g_k P(y_i|\mathbf{x}_i; \mathbf{w}_k^*) .$$

Finally, if $\mathcal{P}(\mathbf{x}) > 0.5$, a lightning-induced outage is predicted. Otherwise, normal operation is assumed.

5. RESULTS

5.1. Experimental Setup

The longitude, latitude and elevation of each substation were used to calculate the distances between them and thus construct the corresponding spatial distance graph. A 128- dimensional spatial embedding was learned for each node (substation) using *node2vec*. The obtained substations' embeddings were concatenated to their corresponding measurement vectors, thus extending the total number of features to 209. The experiments were conducted using different cutoff years for prediction. All models were trained on data from all years prior to 2010, 2014, 2018, and tested on each cutoff year. Upon training, the prediction performance of the Mixture-of-Experts model (MEM) was assessed and compared to several baseline models. Those are listed as follows:

- **Logistic Regression (LR):** A classical logistic regression model that outputs the probability of a binary outcome. The application of logistic regression for outage probability estimation is discussed in [20].
- **Spatially-Aware Logistic Regression (SALR)** [21]**:** A logistic regression variant trained on the substations' spatial embeddings in addition to the original measurements.
- **Subbagging** [22]**:** A subsampling-based ensemble model which considers sampling at random without replacement to generate the training subsets its base models. Each base model in this case is a spatially-aware LR model.
- **Random Partitioning:** A variant of subbagging that utilizes sampling without replacement to generate *disjoint* training subsets for the base SALR models.

5.2. Prediction Performance

Importance of Spatial Information for Event Prediction. The relevance of the spatial information to the task of predicting lightning outages was analyzed by assessing the prediction capability of a single LR model with, and without, using spatial information. The obtained results for the three cutoff years are summarized in Table 2.

From Table 2, it can be observed that SALR outperforms LR at predicting lightning outages. This is consistent across all classification measures. Once spatial information is incorporated, SALR obtains an increase of ~2.3-5.3% in classification accuracy. The gap between the classification performance of SALR and LR increases as more recent cutoff years are used. The underlying reason is most likely related to the larger quantity of training data that becomes available for more recent cutoff years. Furthermore, SALR appear to increase recall by ~5.4-35.7% which is really beneficial, given that recall is of high importance for the task at hand since correctly detecting a certain lightning outage of interest is more important than detecting the majority of the events that do not relate to that outage. Although the largest performance lifts are obtained for precision i.e. the fraction of normal events that were not mispredicted as lightning outages, considerable lifts are also obtained for F1-Score (~3.4-22.2%). This strongly suggests that the spatial information reflected in the substations' spatial embeddings is beneficial for prediction of lightning-induced outages.

Table 2. Prediction performance comparison of LR and SALR across different cutoff years.

Cutoff year	Model	Acc.	AUC	Precision	Recall	F1-Score
2010	LR	0.8445	0.9163	0.7258	0.9296	0.8100
	LR (spatial)	**0.8642**	**0.9242**	**0.7652**	**0.9611**	**0.8373**
2014	LR	0.8289	0.8684	0.7341	0.9383	0.7956
	LR (spatial)	**0.8731**	**0.8795**	**0.8347**	**0.9488**	**0.8565**
2018	LR	0.9085	0.6935	0.5785	0.9356	0.6308
	LR (spatial)	**0.9370**	**0.7576**	**0.7848**	**0.9620**	**0.7709**

Leveraging Multiple Spatially-Aware LR models. Considering the observations from Table 2 indicating that SALR consistently outperforms LR, SALR is further compared to the other alternatives that utilize multiple SALR models. Namely, the predictive performance of the MEM model and the other ensemble-based variants was compared under different parameter settings. For each cutoff year, $K = 1, \dots, 5$ SALR base models were used to train MEM and the ensemble models. No more than $K = 5$ base models were considered since it was observed that the models will not significantly benefit from partitioning the 639 substations into more than 5 regions. In the case of the ensemble variants, for each value of K the base models' subsets were generated using two different randomization strategies: subsampling and disjoint partitioning by sampling without replacement. Since MEM partitions the data into K disjoint subsets, the subsampling fraction for

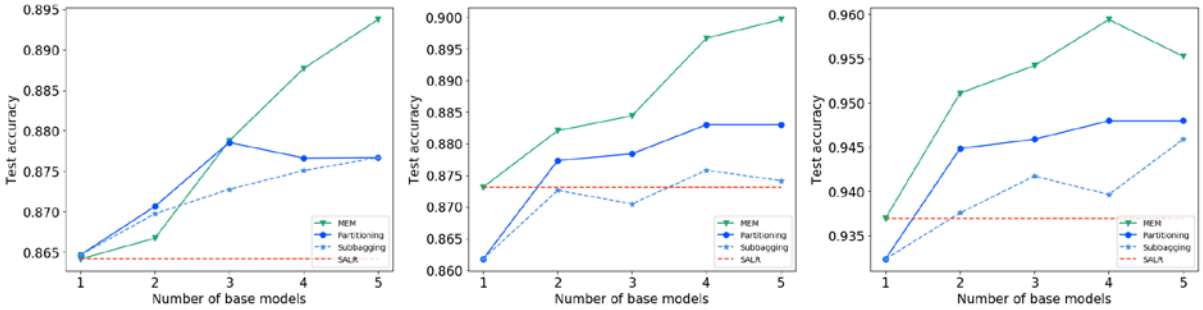


Figure 3. Overall testing accuracy of MEM and alternatives when run with different number of base models in case 2010 (left), 2014 (middle) and 2018 years (right) as cutoff years.

both subsampling and disjoint partitioning was set to $1/K$ assuming uniform training subset sizes for fair comparison. The testing accuracies of all models are presented in Figure 3.

From Figure 3, first it can be observed that MEM and in most cases the ensemble variants outperform a single SALR model for almost all of the parameter settings and gradually achieve greater performance as K increases. The ensemble variants, and especially MEM, achieve greater accuracy when more recent cutoff years. This is most likely due to the fact that more data becomes available for training and thus more patterns become discernible from the data. For the subsampling-based ensembles, random partitioning appears to perform better than the classical subbagging. Nevertheless, MEM generally outperforms both the partitioning and subsampling-based ensemble variants for most of the values of K. In the case when 2010 is used as a cutoff year, random partitioning seems to achieve the highest prediction accuracy until the number of base models reaches 3. Once the substations are partitioned into more than 3 regions, MEM starts to obtained higher accuracy than random partitioning and the other baselines. As for predicting lightning outages for 2014 and 2018, MEM consistently yields the highest prediction accuracy irrespective of the number of base SALRs used. This essentially supports the assumption that training the base models on certain spatial regions (or communities) is more relevant to the task at hand than training them on randomly sampled subsets. There also exists a parameter setting, in all three cases, for which MEM obtains the highest overall accuracy, that being: 0.8937 ($K = 5$), 0.8997 ($K = 5$) and 0.9594 ($K = 4$) when 2010, 2014 and 2018 are used as cutoff years, respectively.

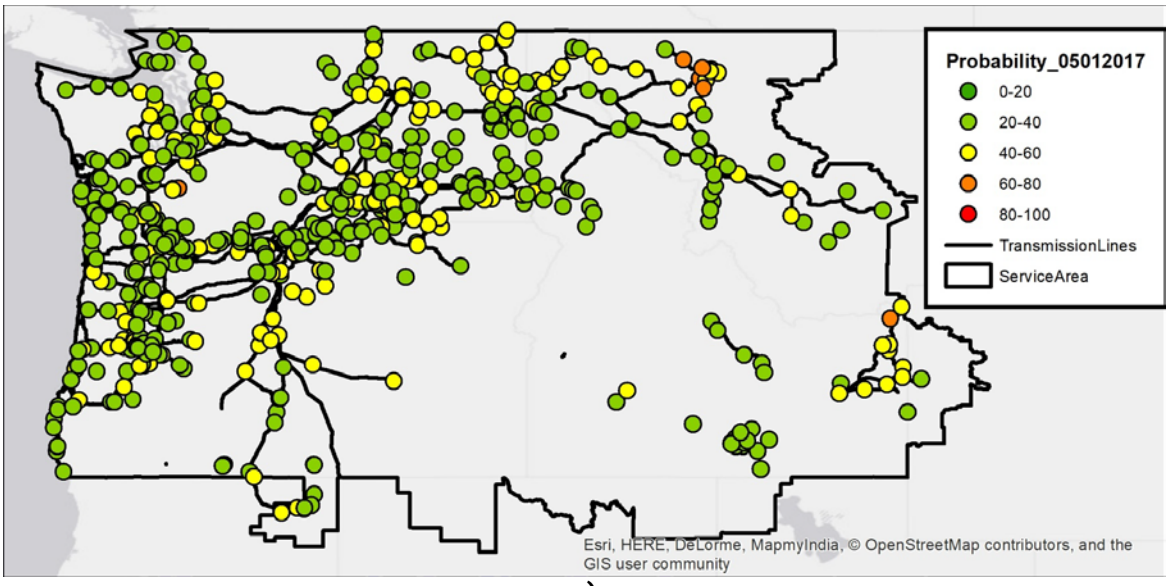
5.3. Results Mapping

The maps with predicted probabilities are presented in Fig. 4 for two cases: a) one hour period on 05/01/2017 that did not have any lightning related outages in the network, and b) one hour period on 05/04/2017 that had multiple lightning caused outages in the network. We can observe from the Fig. 4 that the algorithm is capable of precisely isolating the network area that will be affected by lightning outages by identifying most of the outages with 80% probability or higher, while the rest of the network has lower outage probabilities. Fig. 5 demonstrates the closer look into the outages that occurred on 05/04/2017 and their predicted probabilities.

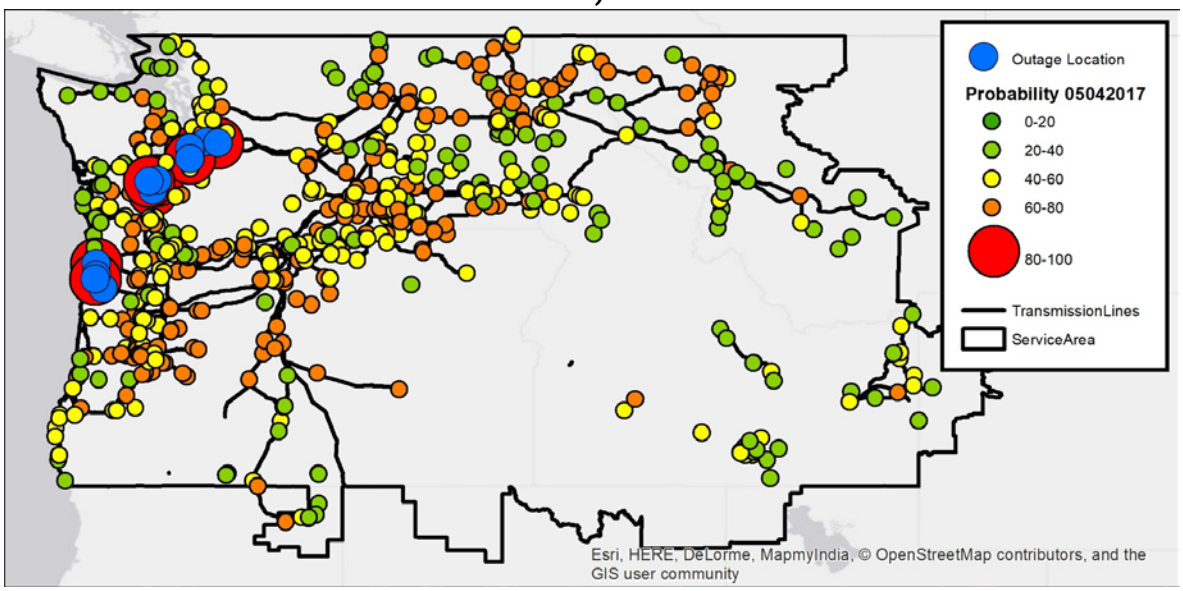
6. CONCLUSIONS

The contributions of this paper are summarized as follows:

- We developed a precise localized prediction of lightning outage probability in the transmission network.
- We used a mixture-of-experts model based on spatially aware logistic regression to provide fast and accurate prediction.
- The results demonstrate accuracy over 80% with Area Under the Curve being larger than 0.75 for all cases with prediction using spatial embedding.



a)



b)

Figure 4. Lightning outage probability maps

- Results demonstrate that the model is capable of isolating the area that will be affected by lightning with very high probability.

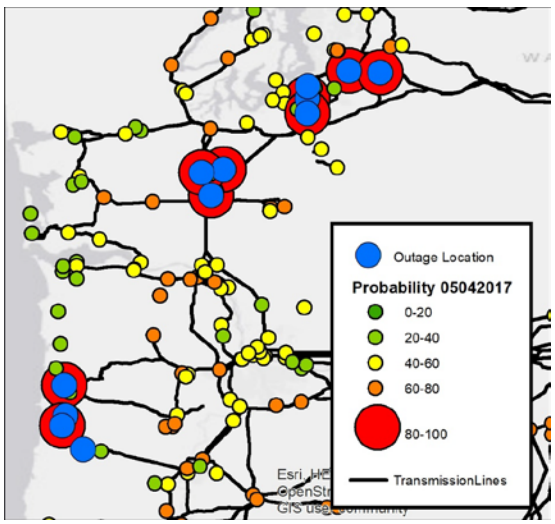


Fig. 5 Closer look into the outages on 05/04/2017

BIBLIOGRAPHY

[1] A. R. Hileman, "Insulation Coordination for Power Systems," CRC Taylor and Francis Group, LLC, 1999.

[2] T. Dokic, et al., "Risk Assessment of a Transmission Line Insulation Breakdowndue to Lightning and Severe Weather," HICCS – Hawaii International Conference on System Science, Kauai, Hawaii, January 2016.

[3] 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.
[4] A. M. Mousa, et al., "Effect of shielding by trees on the frequency of lightning strokes to power lines," IEEE Transaction on Power Delivery, Vol. 3, No. 2, pp. 724-732, April 1988.
[5] T. Sadovic, et al., "Expert System for Transmission Line Lightning Performance Determination", CIGRE Int. Colloq. on Power Quality and Lightning, Sarajevo, Jun. 2012.
[6] A. L. Orille-Fernandez, et al., "Optimization of Surge Arrester's Location," Power Delivery, IEEE Transactions on, Vol. 19, No, 1, pp. 145-150, January 2004.
[7] R. Shariatinasab, et al., "Probabilistic Evaluation of Optimal Location of Surge Arresters on EHV and UHV Networks Due to Switching and Lightning Surges," Power delivery, IEEE Transactions on, Vol. 24, No. 4, pp. 1903-1911, October 2009.
[8] E. Bullich-Massague, et al., "Optimization of Surge Arresters Location in Overhead Distribution Networks," Power Delivery, IEEE Trans. on, Vol. 30, No. 2, pp. 674-683, April 2015.
[9] M. Kezunovic, et al., "Optimal Placement of Line Surge Arresters Based on Predictive Risk Framework Using Spatiotemporally Correlated Big Data," CIGRE General Session, Paris, France, Aug. 2018.
[10] I. M. Rawi, et al., "Lightning study and experience on the first 500kV transmission line arrester in Malaysia," International Conference on Lightning Protection (ICLP), Shanghai, China, 2014.
[11] W. Sones, S. M. Wong, "Overview on Transient Overvoltages and Insulation Design For a High Voltage Transmission System," High Voltage Engineering and Application (ICHVE), 2010 International Conference on, New Orleans, LA, 2010.
[12] BPA, "Outage and reliability reports," [Online] Available: https://transmission.bpa.gov/Business/Operations/Outages/
[13] Vaisala, "National Lightning Detection Network NLDN" [Online] Available: https://www.vaisala.com/en/products/data-subscriptions-and-reports/data-sets/nldn
[14] 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
[15] NOAA, "NDFD Data and Support" [Online] Available: https://www.weather.gov/mdl/ndfd_home
[16] Google, Google Maps Platform, Elevation API, [Online] Available: https://developers.google.com/maps/documentation/elevation/start
[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, August 2016.
[18] T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality," Advances in neural information processing systems (pp. 3111–3119), 2013.
[19] K. Murphy, "Logistic regression," Machine Learning: A Probabilistic Perspective, Chapter 8, pp. 245– 279, 2012.
[20] M. Kezunovic, et al., "Systematic Framework for Integration of Weather Data into Prediction Models for the Electric Grid Outage and Asset Management Applications," The Hawaii International Conference on System Sciences – HICSS, Waikoloa Village, Hawaii, January 2018.
[21] T. Dokic, et al., "Spatially Aware Ensemble-Based Learning to Predict Weather-Related Outages in Transmission," The Hawaii International Conference on System Sciences – HICSS, Maui, Hawaii, January 2019
[22] S. Andonova, et al., "A simple algorithm for learning stable machines," Proceedings of the 15th European Conference on Artificial Intelligence, pp. 513–517, July 2002.