Combining Remotely Sensed Imagery with Survival Models for Outage Risk Estimation of the Power Grid

Arpit Jain *
GE Global Research
San Ramon, CA, USA
arpitjain1@gmail.com

Tapan Shah *
GE Global Research
San Ramon, CA, USA
tapan.shah@ge.com

Mohammed Yousefhussien *
GE Global Research
Niskayuna, NY, USA
myhussien@gmail.com

Achalesh Pandey
GE Global Research
San Ramon, CA, USA
achalesh.pandey@ge.com

Abstract

Vegetation management of power grids is essential for reliable distribution of services, prevention of forest fires and disruption of electricity due to tree fall. In this paper, we introduce a vegetation analysis system that utilizes information from GIS data, aerial and satellite imagery to estimate vegetation profile within a buffer zone. This vegetation profile is further combined with operational parameters of the grid to develop a survival model which predicts the outage risk of a power-line in an electrical grid. Using historical data, we show that the risk scores thus obtained are significantly better at developing trimming schedules for grid power-lines, compared to existing available methods.

1. Introduction

Aerial and satellite image analysis is becoming increasingly important in addressing the challenges related to disaster management, infrastructure and urban planning, Transmission and Distribution (T&D) maintenance. Utility companies are under immense pressure to maintain their frail T&D infrastructure with limited resources. In this paper, we focus on solving vegetation management problem which is critical to preventing catastrophic events such as forest fires and large scale outages due to vegetation encroachment.

Utility companies perform two types of tree trimming activities on power-lines to address vegetation encroachment: (a) scheduled or planned trimming (b) hot-spot or unplanned trimming. Scheduled trimming is performed based on a schedule designed by inspector based on operation knowledge or ad-hoc combination of operations variables such as number of customer impacted or number of outages experienced on a given power-line without the knowledge of actual state of vegetation on the ground. Hot-spot trimming is performed on-need basis when customers complain about encroachment or due to adverse weather event.

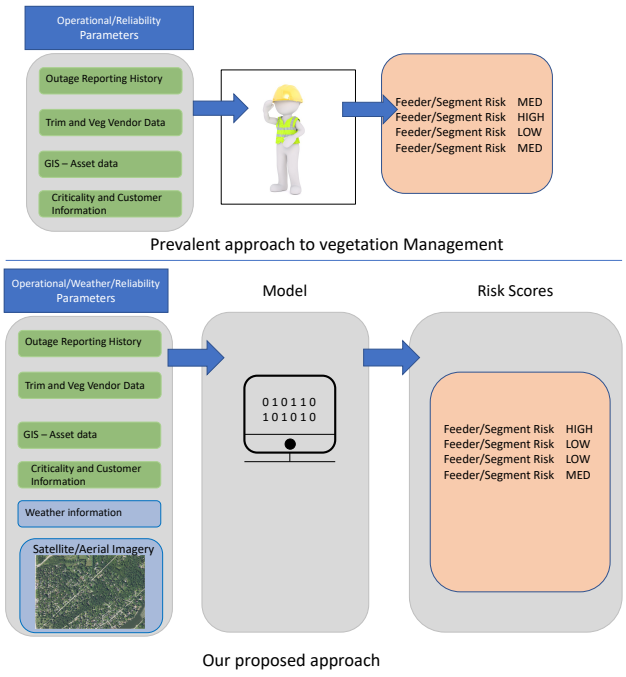

Figure 1. An illustration comparing the prevalent and our proposed approach to vegetation management.

Currently, a vegetation manager determines the risk associated with each power-line and creates a trimming sched-

*Equal Contribution

ule based on historical outages and operational data [26]. This approach is subjective, biased and fails to incorporate the actual vegetation encroachment on the ground, thereby resulting in a sub-optimal trimming schedule which further leads to higher vegetation related outages as well as higher maintenance cost [10]. Figure 1 shows the current approach to vegetation management and the motivation for our proposed approach.

In recent times, utility companies have started investigating the adoption of LiDAR-based vegetation management [11, 26]. While LiDAR provides high accuracy and fidelity for vegetation encroachment, they are still not widely adopted due to high cost associated with data collection and processing. On the other hand, satellite and aerial imagery provides a compelling solution with the benefit of large-scale coverage at lower cost. Even though our proposed approach leverages aerial and satellite imagery in this paper, it can easily incorporate LiDAR-based vegetation information when available.

In this paper, we propose a machine learning framework (Figure 1) to address tree trimming problem using survival modeling which combines GIS data, aerial and satellite imagery, and other operational parameters to estimate the outage risk which can be used to derive an optimal scheduling. The same framework can also address the issue of spot trimming to identify the most risky power-lines which may be in need of trimming prior to their scheduled trim time.

Vegetation segmentation in large scale aerial and satellite imagery can be challenging due to the lack of annotated data. In the absence of large budget for annotation task, we propose a two-stage approach to training the CNN model for vegetation modeling using semi supervised strategy. Our experiments shows that with just a few number of annotated samples, we can train a CNN vegetation segmentation model which generalizes well for the entire local grid territory. Another challenge related to vegetation segmentation in freely available aerial imagery data (NAIP dataset - 1m resolution) is the low resolution of trees. In the NAIP dataset, a typical tree width spans a range of 10-35m on the ground which translates to 10-35 pixels, and twice that value for WorldView-3 satellite at 0.5m resolution. We propose a modified segmentation approach based on [40] to address the issues of "small" tree pixels by avoiding pooling operations in encoding layers.

Risk associated with vegetation-related outages on a power-line is also dependent on several other operational, environmental and reliability factors besides tree encroachment. For example, vegetation encroachment on a short power-line serving 20,000 customer is more important to trim earlier than a long power-line serving 10 customers. Certain power grids are more prone to vegetation-related outages due to weather conditions than others. To model the risk associated with each power-line, machine learning approach will have to incorporate these additional variables in prioritising trimming. We propose a survival modeling based generic framework to model this requirement. Survival modeling can incorporate weather, operational Key Performance Indicators (KPIs) and other reliability factors in conjunction with vegetation encroachment from aerial and satellite imagery to provide a holistic outage risk score combining these factors.

In this paper, our main contributions are as follows:

- We combine aerial, satellite, and GIS data to define tree encroachment around power-line segments for a large-scale analysis. Our vegetation analysis can be done at multiple scales (i.e. power-line segment, a feeder, or a full circuit level)
- We utilize semi-supervised approach to reduce annotation burden. We utilize weak classifier to generate pseudo labels which are then fed into segmentation network for learning.
- We use *Cox proportional hazard* survival modeling technique to combine the vegetation KPIs with other reliability and operational attributes of a power-line as well as the prior trimming schedule to assign a risk to quantify the likelihood of a future outage of a power-line.

1.1. Prior Work

Recently, aerial and satellite imagery are gaining an increased attention from various research communities for solving real world problems such as disaster response, urban planning, precision agriculture, and autonomous driving [12]. With increasing blackouts due to to vegetation encroachments for power-lines, it is necessary to maintain clearance within the right-of-way (ROW) buffer zone to ensure uninterrupted distribution of electricity. Therefore, utility companies are exploring the use of such rich data to improve reliability and reduce cost of maintaining grid assets [22].

Researchers have investigated the task of vegetation segmentation and tree encroachments from various sensing modalities. For a comprehensive overview of the modalities and algorithms developed for vegetation segmentation and tree encroachment before the proliferation of deep neural networks, we refer the reader to following survey papers [28] [2]. Convolutional Neural Networks have shown tremendous success in semantic segmentation tasks and a lot of recent works have focused around using deep neural networks for vegetation segmentation. [23] proposed a multi-task learning framework for pan-sharpening and semantic segmentation of trees in a 25 km^2 0.3m resolution WorldView-3 satellite data. [25, 27, 32, 35] proposed several

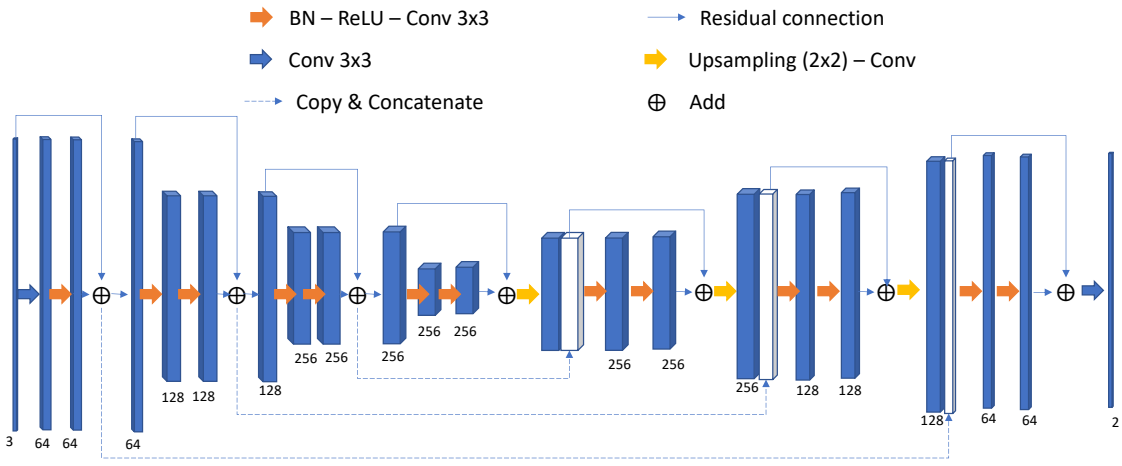

Figure 2. Residual U-Net architecture used for vegetation segmentation

CNN based architectures for land cover segmentation which can be extended to vegetation segmentation. In [33], a deep learning-based detection framework that utilizes the images obtained from vision sensors mounted on power transmission towers was proposed. The method includes three cascaded modules: detection of vegetation regions using Faster R-CNN, detection of power lines based on the Hough transform, and detection of vegetation encroachment using depth from stereo. While the method was able to detect powerlines and tree encroachment, it relies on ground-based sensors close to the area of interest which hinders the ability to cover large scale area for encroachment analysis.

A large scale study, similar to ours in terms of scale, was conducted in [3] where the authors performed tree-cover delineations using NAIP imagery. The study used 150 handcrafted statistical features computed from super-pixel regions found using region growing algorithm. The features per pixel are then classified using a fully-connected feed forward neural network. A conditional Random Fields (CRF) step is then applied to incorporate contexts for region consistency. While their work spanned a large scale area, they used handcrafted features which may not generalize well.

A lot of the prior work have focused on vegetation related risk modeling in the context of adverse weather. [38] used LiDAR to identify high risk trees and combined prior trim and infrastructure data with vegetation related information using random forest to identify high risk power lines in the event of a storm. Similarly, [6] proposes a outages prediction model using regression trees based on weather prediction outputs, soil information, vegetation, electric utility assets, and historical power outage data. However, none of these approaches address the issue of vegetation management from trimming perspective.

Survival models for risk modeling has been used for several decades in multiple applications like cancer survival [7, 39], predictive maintenance and remaining useful life of industrial systems [34, 36] but our work is among the first to apply it to vegetation trimming problem.

[15] studied the vegetation related outages under two categories: growth-related and weather-related outages. Two types of models for the two tasks, namely Time series and nonlinear machine learning regression models, are proposed to conduct the outage prediction using vegetation, prior outages, weather and geographical data. The work most relevant to ours is [13] where a Gaussian conditional random field approach was proposed to combine operational, weather and vegetation related parameters for optimal tree trimming. The tree segmentation was obtained through super-pixel clustering and assignment approach. However, to the best of our knowledge, there is no prior work which combines aerial and satellite imagery with advanced deep neural network-based computer vision techniques and survival models to develop an end-to-end vegetation management system.

2. Vegetation Analysis

In this section, we will discuss in more details our approach for vegetation segmentation in aerial and satellite imagery.

2.1. Vegetation Segmentation

We treat this problem as a two class segmentation problem as our focus is solely on vegetation encroachment. However, proposed approach can be extended to multiple classes. Several semantic segmentation architectures have been proposed in literature. We use the Residual U-Net architecture from [40] for vegetation segmentation. The network is shown to work well for road segmentation in aerial imagery and parameters work well for semantic classes with

small footprints such as individual trees. We have minor modifications to the network, found empirically, in number of filters but the architecture is shared with the reference. Figure 2 shows the network architecture of our proposed approach. We use equally weighted binary cross-entropy and

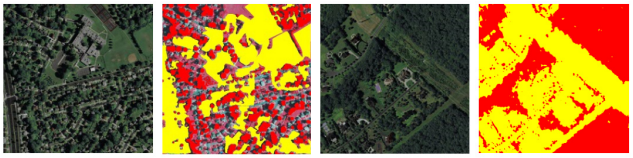

Figure 3. Result of our data annotation strategy(a) Sample Train data (b) Sample Manual annotation (c) Sample Unlabeled data (d) Sample Pseudo label

dice loss as the loss function to minimize. Additional details related to training is given in the experiment section. In the next section, we will discuss the semi-supervised learning approach to reduce the annotation requirement and improve networks robustness against unseen territories.

Image annotation is a tedious and expensive task. Given, we are performing vegetation segmentation on a large territory, it is hard to annotate different variety of vegetation across the territory. In order to ameliorate the lack of large annotation dataset, we leverage a two-stage learning strategy with a human in the loop to annotate our aerial and satellite imagery. Initially, we manually label a small set of images. In our case, we labeled 60 images of 512×512 pixels. Then, using this small set of labeled images, we use a random forest [4] to introduce two sources of randomness to help control over-fitting. The first is *bagging*, where each tree is grown using a bootstrap sample of training data, and the second one is *random feature splitting* where the best split at each node is chosen from a random sample of features instead of all. In our case, we trained 200 decision trees estimators with a batch size of 100 pixels with unlimited depth. We consider two random features at each node. To train the model, we used a set of hand-crafted features such as Gabor filters, mean, min , max, and the variance at each pixel with 5 neighborhood window. After training the random forest on the manually labeled data, we then use this model to automatically annotate a larger set of 2k images that are then used to train our deep learning along with the manually labeled data. Figure 3 shows a manually annotated image and the prediction of the random forest classifier.

2.2. Vegetation KPIs

In order to estimate outage risk associated with vegetation, we need to estimate certain KPIs like area of vegetation encroachment, length of encroachment, etc. with respect to the utility grid and structures. These KPIs will then feed to our machine learning model to estimate outage risk along with other operational, environmental and reliability parameters. Usually, 5m distance on either side of the power line is considered no encroachment or ROW zone. Any vegetation encroaching within the right-of-way is considered unsafe and needs to be trimmed. We first perform tree segmentation on all the tiles extracted from the aerial imagery for the territory. We then combine all the tiles together to create a single vegetation probability map for the whole territory. The GIS information is then overlayed on the probability map to define right of way by placing a 5m buffer on either side of the power line. The power-line itself is composed of a set of line segments. For each line segment, we can compute vegetation area and length of vegetation encroachment with an oriented bounding box defining the right-of-way zone. These Key Performance Indicators (KPIs) at individual section of a power-line represents the segment-level vegetation risk. Inspector can use this information to identify regions where there are significant encroachment occurring within a power line. This information together with other variables can help identify future potential outages associated with each power-line and plan the trimming optimally. In the next section, we will describe our survival modeling based approach which can seamlessly combine vegetation related KPIs with other variables for risk estimation.

Figure 4 shows an overview of our vegetation segmentation on a sample patch from NAIP dataset and the corresponding vegetation KPIs.

3. Risk Modeling Using Survival Models

In this section, we introduce survival model as a risk estimation methodology and illustrate how it can be used to model outage risk of power-lines as a function of vegetation KPI and other attributes.

3.1. Survival Models: Basic principles

Survival models are a class of statistical models which obtains estimates of time of a particular event of interest (like death, failure, outage etc.) i.e. survival times. The key property of a survival model is the ability to incorporate censored data i.e. if an individual does not have an event till the observation time, they are described as censored. It means that after the data is collected, the individual may or may not have an event. This special property makes survival models more appealing to estimate event times as compared to standard regression based methods (where we would have to drop the censored data).

Some definitions related to survival modeling are as follows [24, 37]:

- **Survival function** is the probability that the event of interest occurred after specified time t.

$$S(t) = P(T > t),$$

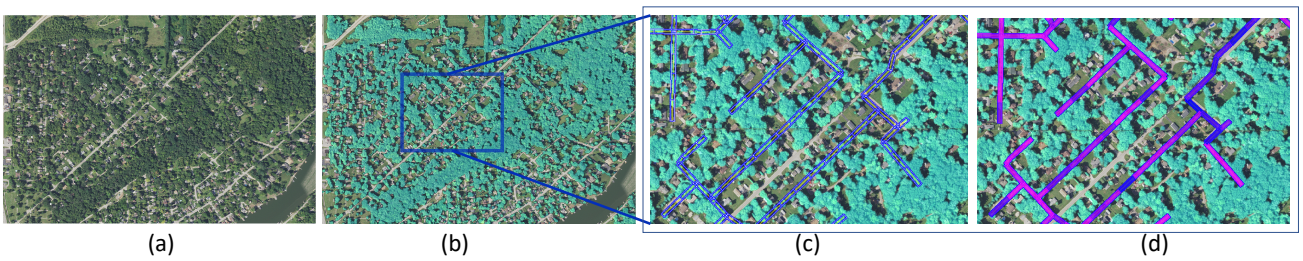

Figure 4. Overview of our vegetation segmentation and KPI estimation. (a) an image tile from NAIP aerial imagery. (b) shows the result of our vegetation segmentation overlaid on the image. (c) a zoomed-in region showing the power lines along with right-of-way non-encroachment buffer zone (defined by blue color lines). (d) Shows each power-line segment color coded from blue to pink in proportion to the the area of vegetation encroachment. More pink indicates more vegetation overlap and more blue indicates the opposite.

where T is the survival time.

- **Cumulative death distribution** is $F(t) = 1 - S(t)$ and the **death density function** is $f(t) = \frac{dF(t)}{dt}$
- **Hazard function** is the likelihood of the event occurring at time t, given no event occurred before t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T \leq t + \Delta t | T \geq t\right)}{\Delta t} \quad (1)$$

$$= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} \quad (2)$$

$$= \frac{f(t)}{S(t)} \quad (3)$$

Concretely, each data-point in the data-set used to develop survival models is a tuple (X, y, δ) where X is the feature vector (also called as covariates), δ is an indicator variable which is 1 for uncensored instance and 0 for a censored instance, and y is the time of event (survival time or failure time in some literature) for uncensored instances or the observation time for censored instances. This is unlike a supervised regression or classification problem where each data-point contains a tuple of 2 points. In addition to specialized models like Cox Proportional Hazard model [5], several machine learning methods like logistic regression [39], support vector machines [30], random forests [19], neural networks [16, 21] have been modified to obtain a transfer function which estimates the survival function or hazard function as a function of the feature vector [1]

3.2. Cox Proportional Hazards (PH) Model

Cox PH model is a semi-parametric model which models the hazard function $h(t, X)$ as a product of the time component $\lambda_0(t)$ and the feature component $\eta(X)$

$$h(t, x) = \lambda_0(t)\eta(x), \quad (4)$$

where

- $\lambda_0(t)$ is the unspecified baseline function and,
- $\eta(x)$ is the risk due to the features modeled as $\exp\left(\sum w_i X_i\right)$.

Note that we have defined the hazard function as a function of the feature vector X. The key point is that there is a natural way in which the hazard changes with time, which is the same for any two individuals. The difference in the risk is due to the features associated with each individual. The parameters $\{w_i\}$'s are estimated by maximizing the partial likelihood using the Newton Raphson's method.

3.3. Connecting Survival Models and Risk Modeling of Power Lines

The key goal of our vegetation management system is to develop a multi-variate model which can estimate the risk of each power-line. The 3 types of attributes available for each power-line to develop the risk model are explained:

1. **Reliability** attributes [1] like customers impacted (CI), number of outages (NO), customer average interruption duration index (CAIDI), system average interruption duration index (SAIDI) and system average interruption frequency index (SAIFI), of previous year, given an indication of risk for the current year.
2. **Environmental** attributes like vegetation KPIs (as described in Section 2) and weather attributes like wind gust, precipitation and temperature [8].
3. **Operational** attributes like number of customers, overhead coverage of the power-line and time since the last trim.

Assumptions

Before explaining the modeling methodology, we highlight two key assumptions below.

1. The first of January for each year is the instant of observation and decision-making (i.e. all data prior to

[1] There are non-parametric methods like Kaplan-Meir estimate [20] which do-not take into account the feature vector.

that day can be used to estimate the future outage risk of each power-line for that year).

2. Each power-line is trimmed every 4 years. We assume the risk is reset after each trim i.e. the power-line is "reborn". Also, the maximum life of a power-line is 4 years [2].

In the next discussion, we show how the data collected from power lines fits naturally to the requirements of survival model paradigm. For any power-line (F) on the 1^{st} of January of any year (t), the training data tuple consists of $(X_F(t-1), y_F(t), \delta_F(t))$ where $X(t-1)$ is the vector of reliability, environmental and operational values from previous years, $\delta_F(t)$ is an indicator which is 1 if an outage occurred for that power-line in the year t and $y_F(t)$ is the number of days between the last trim and outage, if an outage occurred else it is the number of days between the last trim and end of the year or the trim date. We observe that the data is in a format similar to the one required to train survival models (explained in Section 3.1). Collecting similar data for multiple power-lines over the training period, we estimate the parameters of a Cox proportional hazard model using an implementation provided in [17]. The key reason for preferring the Cox proportional hazard model is the easy interpretability: the contribution of the attributes to the "hazard" can be easily obtained from the magnitude of the estimated coefficients $\hat{w}_i$ in (4). Additionally, the factorization of the baseline hazard and the feature contribution makes it trivial to compute a time-independent risk score ($\sum_i \hat{w}_i X_i$) (it is non-trivial for other methods).

4. Experimental Results

4.1. Dataset

Our study area covers $1760mi^2$ within two states in the US. We collected 100 georeferenced NAIP and Worldview-3 tiles where each tile covers a 3.7×4.66 sq. miles. NAIP imagery consists of 4 bands and is acquired at a 1m ground sample distance (GSD), while the Worldview-3 imagery consists of 8 multi-spectral bands along with SWIR and CAVIS bands with GSD ranging from 0.3m to 10m. From both sources, we used only 4 bands, namely Red, Green, Blue, and Near-IR. Along with the imagery data, we also utilized georeferenced GIS data indicating the location and the extent of each power-line in the area of interest. Due to the proprietary nature of the GIS data, the location is, therefore, undisclosed.

[2] A natural question which can be raised is that why can't we make the prediction every month (or every day)? Even though our methodology can be naturally extended, we provide a yearly risk for 2 main reasons. Firstly, from a decision-making point of view, the vegetation inspector will want a list of high risk power-lines at the beginning of the year to plan the maintenance schedules. Secondly, it would require getting the images at a higher temporal frequency which can be expensive.

4.2. Performance of Vegetation Segmentation

We trained our segmentation model using a small set of 1860 images. However, we increased the training data by augmenting using various strategies such as flipping, rotating, adjusting brightness and contrast, as well as saturation adjustment. Our model was trained for 50 epochs with a batch size of 8, and a learning rate of 0.0001 using Adam optimizer. To reduce the effect of over-fitting, we applied polyak averaging using exponential decay of the last 10 checkpoints, and applying label smoothing to the ground truth data. We evaluated the performance of our segmentation approach using two separate metrics. We evaluated the performance of our segmentation approach on a validation set of 200 images. We achieved a mIoU score of 97.2%. We also evaluated the performance of our segmentation in ROW zone against a proprietary LiDAR-based ground truth tree masks. Our approach obtained a mIoU score of 86% against LiDAR data. This shows that our approach can provide comparable results w.r.t LiDAR data for tree segmentation within the buffer zone.

4.3. Evaluation of Risk Modeling

4.3.1 Data Preparation

In order to validate the performance of the risk modeling method, we use the data from a utility company on a population of 2000 power-lines located in a single geographical area. The reliability feature vector used were a weighted sum of CI, number of outages (NO) and SAIDI for the previous 2 years where the weight for the previous year was 2/3 and the weight for the year before was 1/3. For example, for 2019, the reliability features of a power-line is $[\frac{2}{3}CI(2018) + \frac{1}{3}CI(2017), \frac{2}{3}NO(2018) + \frac{1}{3}NO(2017), \frac{2}{3}SAIDI(2018) + \frac{1}{3}SAIDI(2017)]$. This was appended with vegetation KPIs computed in Section 2 as well as the overhead mileage of each power-line. The temporal component required for the survival modeling is computed as discussed in Section 3.3.

The data between 2016 and 2018 is used to train the model whereas data in 2019 is used to evaluate the performance of our model.

4.3.2 Evaluation

Discriminate measures are typically used to to validate the performance of risk models [9, 29, 31] models. These measures capture how well can the model discriminate between the "low" and "high" risk entities.

Ground truth: low and high risk: Identifying *actual* high and low risk power-lines (pl) is a subjective business decision. Based on expert feedback, two definitions of identifying high and low risk power-lines for a given year were decided

- **DEF 1**: If CI>120 then the power-line is high risk else if CI<5 it is low risk, else they are removed from evaluation.
- **DEF 2**: If NO>2 then the power-line is high risk else if NO<1 it is low risk, else they are removed from evaluation.

A key thing to note is that this are retrospective definitions and can only be used to conduct evaluation on historical data.

Baseline: Previous Year CI The business historically used previous years' CI as a risk measure R_{CI} i.e. to forecast and schedule planned and unplanned maintenance for 2019 on 1 January 2019, corresponding value of CI from 2018 is used as a risk measure. Thus if for a power-line, $R_{CI}(2019) = CI(2018) > \tau_{CI}$ then the power-line is predicted to be *high risk* for 2019, else it is predicted to be low risk.

Our Method: Survival Risk score We used the risk score R_{su} computed by the survival model i.e. on 1 January 2019, we use the data from previous years to compute risk $R_{su}(2019)$ and if $R_{su}(2019) > \tau_{su}$, the power-line is predicted to be *high risk* for 2019. The values of τ_{su} and τ_{CI} can be chosen according to business considerations based on the precision-recall (PR) or receiver operating characteristic curve. (ROC) [18].

Precision, Recall, False Positive Rate (FPR) an True Positive Rate (TPR): Adapting the standard definitions from a confusion matrix, we define

$$recall = \frac{\text{\# pl predicted \& actually high risk in 2019}}{\text{\# pl actually high risk in 2019}},$$

$$precision = \frac{\text{\# pl predicted \& actually high risk in 2019}}{\text{\# pl predicted high risk in 2019}},$$

$$FPR = \frac{\text{\# pl predicted \& not high risk in 2019}}{\text{\# pl actually not high risk in 2019}}.$$

In Figure 5, we compare the precision-recall curves as well receiver operating characteristic curve (ROC) [18] for DEF1 by varying τ_{su} and τ_{CI} . Similar results are also observed for DEF2.

4.3.3 Analysis of Survival Model

In this section, we analyze the fitted survival model in some detail.

Training Loss: In Figure 6, we shows the evolution of the loss (negative log partial likelihood) and the gradients while fitting the model parameters using Newton Raphson method.

Risk Groups: Using a suitable threshold on the risk score, we divide the power-lines into two groups,"high risk" and "low risk". The distribution of the scores as well as the high and low risk groups are shown in Figure 7.

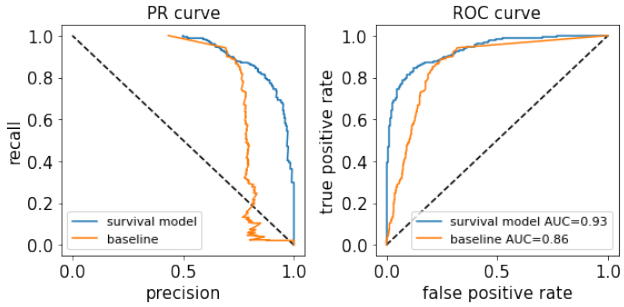

Figure 5. PR and ROC curves comparison between the baseline and survival model based risk scoring.

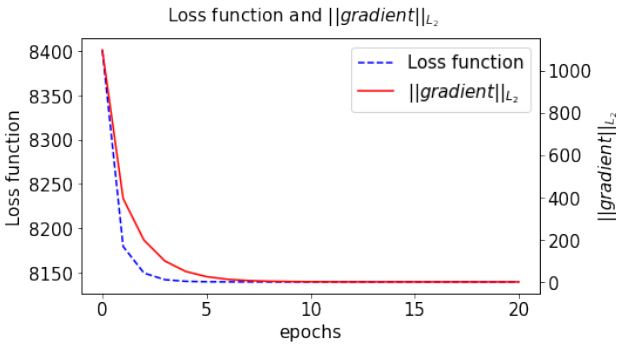

Figure 6. Evolution of training loss and the gradient.

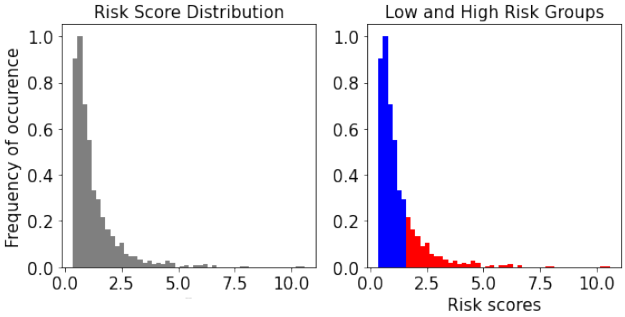

Figure 7. Distribution of risk scores and classifying them into high risk and risk.

Individual Predictions: In Figure 9, we observe that for the power-line which we predicted high risk, the survival function drops sharply and it suffers an outage within 7 months. Conversely, for the power-line which we predicted to be low risk, the survival function drops slowly and there was not outage till the time of recording the data (when that power-line was 30 months since the last trim).

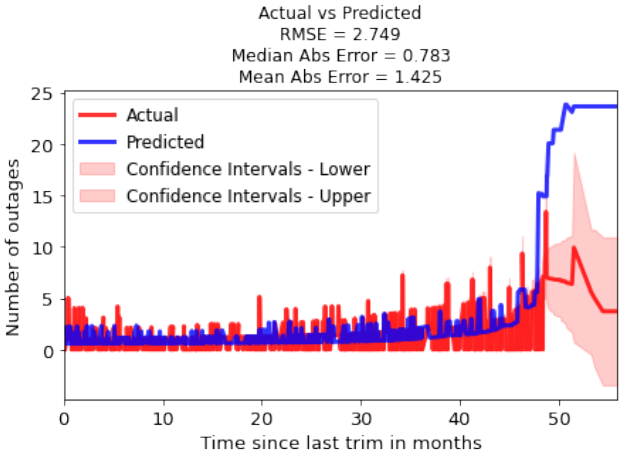

Figure 8. Actual vs Predicted- Number of power-lines experiencing an outage as a function of months since last trim.

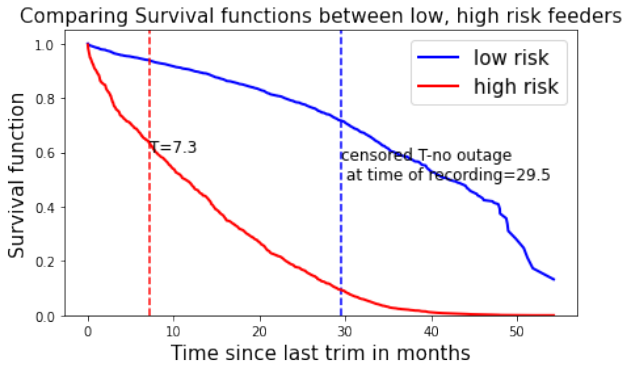

Figure 9. Individual survival functions for 2 power-lines: one which was predicted risk and other which was predicted low risk. The vertical line is the time (since last trim) of the outage or the censored time.

4.3.4 Discussion

In this section, we discuss certain key insights and learning from the evaluation of our risk model.

Performance Improvement: We clearly observe from the results in Figure 5 the numeric gains of our method over baseline methods. The overall ROC-area under the curve is superior by 7%. Additionally, if the business requires a minimum recall of 75%, the survival model based risk can achieve a precision of 90% compared to a 80% precision achieved using the baseline method.

Survival Model vs Classification/Regression Problem: An obvious question arising is "Why cannot we use classification/regression methodology?". This is a valid question and we provide the following reasons on why survival models make a better choice.

- A classification problem requires a *a-priori* labeling of high-risk and low-risk power-lines. Thus whenever business dictates a change of definition depending on the geographical location or year, the model needs to be re-trained. An interesting discussion on the similar use of survival models over classification for cancer prognosis is discussed in [7].

- Another obvious choice is a regression on the time to event. However, this leads to issues on censored data i.e. those power-lines which at the time of data collection have not had any outage. As explained in Section 3.1, survival models provides a unique method to incorporate censored data.

- From a philosophical sense, this agrees with our intuition on how power-lines suffer outage: The factorization of the Cox PH model combines a natural *time varying* deterioration in power-lines (increase in risk) with risk inducing external factors to estimate an overall risk.

5. Conclusion and Future Work

In this paper, we propose a novel vegetation management system which combines information from GIS data, aerial and satellite imagery to estimate vegetation profile, which is further combined with operational and reliability attributes of a power-line using a survival model to estimate a risk score. We envisage 2 main applications of this risk score. First, provide a list of power-lines in decreasing order of risk which can then be combined with spatial clustering to identify optimal *s*cheduled trimming schedules. This is necessary as it involves significant resource cost to move around equipment for trimming. Hence the trimming team would ideally like to first cover a geographical area which has a high proportion of high risk power-lines. Second, provide a list of very high risk power-lines to for *u*nscheduled/spot trimming. This allows to reduce the number of outages and thereby reduce the number of customers impacted.

Often, vegetation related outages are caused not only due to growth but also due extreme weather events [14]. Integrating weather forecasts is an important future line of work to develop a comprehensive vegetation management which can provided combined risk score using both weather forecasts and vegetation density. Another important line of work is to compare the Cox PH method with advanced deep neural network based survival models like [16, 21].

References

[1] IEEE Guide for Electric Power Distribution Reliability Indices. *IEEE Std 1366-2012 (Revision of IEEE Std 1366-2003)*, pages 1–43, May 2012. Conference Name: IEEE Std 1366-2012 (Revision of IEEE Std 1366-2003). 4325

[2] Junaid Ahmad, Aamir Saeed Malik, Likun Xia, and Nadia Ashikin. Vegetation encroachment monitoring for transmission lines right-of-ways: A survey. *Electric Power Systems Research*, 95:339–352, 2013. 4322

[3] Saikat Basu, Sangram Ganguly, Ramakrishna R Nemani, Supratik Mukhopadhyay, Gong Zhang, Cristina Milesi, Andrew Michaelis, Petr Votava, Ralph Dubayah, Laura Duncanson, et al. A semiautomated probabilistic framework for tree-cover delineation from 1-m naip imagery using a high-performance computing architecture. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5690–5708, 2015. 4323

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 4324

[5] N. E. Breslow. Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45–57, 1975. Publisher: [Wiley, International Statistical Institute (ISI)]. 4325

[6] D. Cerrai, D. W. Wanik, M. A. E. Bhuiyan, X. Zhang, J. Yang, M. E. B. Frediani, and E. N. Anagnostou. Predicting storm outages through new representations of weather and vegetation. *IEEE Access*, 7:29639–29654, 2019. 4323

[7] Hung-Chia Chen, Ralph L. Kodell, Kuang Fu Cheng, and James J. Chen. Assessment of performance of survival prediction models for cancer prognosis. *BMC Medical Research Methodology*, 12(1):102, July 2012. 4323, 4328

[8] P. Chen, T. Dokic, N. Stokes, D. W. Goldberg, and M. Kezunovic. Predicting weather-associated impacts in outage management utilizing the gis framework. In *2015 IEEE PES Innovative Smart Grid Technologies Latin America (ISGT LATAM)*, pages 417–422, 2015. 4325

[9] R. B. D'Agostino and Byung-Ho Nam. Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures. In *Handbook of Statistics*, volume 23 of *Advances in Survival Analysis*, pages 1–25. Elsevier, Jan. 2003. 4326

[10] Ashiss Kumar Dash. Vegetation Management: Artificial Intelligence to Preempt Forest Fires. https://www.tdworld.com/vegetation-management/article/20973359/vegetation-management-artificial-intelligence-to-preempt-forest-fires, Nov 11, 2019. 4322

[11] Nick Day. LiDAR for Distribution Vegetation Management. https://www.tdworld.com/vegetation-management/article/20972795/lidar-for-distribution-vegetation-management, Jul 01, 2019. 4322

[12] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 4322

[13] T. Dokic and M. Kezunovic. Predictive risk management for dynamic tree trimming scheduling for distribution networks. *IEEE Transactions on Smart Grid*, 10(5):4776–4785, 2019. 4323

[14] Milad Doostan, Reza Sohrabi, and Badrul Chowdhury. A Data-Driven Approach for Predicting Vegetation-Related Outages in Power Distribution Systems. *arXiv:1807.06180 [cs, stat]*, Mar. 2019. arXiv: 1807.06180. 4328

[15] Milad Doostan, Reza Sohrabi, and Badrul Chowdhury. A data-driven approach for predicting vegetation-related outages in power distribution systems. *International Transactions on Electrical Energy Systems*, 30(1):e12154, 2020. e12154 ITEES-19-0274.R1. 4323

[16] Stephane Fotso. Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework. *arXiv:1801.05512 [cs, stat]*, Jan. 2018. arXiv: 1801.05512. 4325, 4328

[17] Stephane Fotso et al. PySurvival: Open source package for survival analysis modeling, 2019–. 4326

[18] Patrick J. Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, Mar. 2005. 4327

[19] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, Sept. 2008. arXiv: 0811.1645. 4325

[20] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958. 4325

[21] Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, 18(1):24, Dec. 2018. arXiv: 1606.00931. 4325, 4328

[22] Mladen Kezunovic, Pierre Pinson, Zoran Obradovic, Santiago Grijalva, Tao Hong, and Ricardo Bessa. Big data analytics for future electricity grids. *Electric Power Systems Research*, 189:106788, 2020. 4322

[23] Andrew Khalel, Onur Tasar, Guillaume Charpiat, and Yuliya Tarabalka. Multi-task deep learning for satellite image pansharpening and segmentation. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 4869–4872. IEEE, 2019. 4322

[24] David G Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 2010. 4324

[25] Tzu-Sheng Kuo, Keng-Sen Tseng, Jia-Wei Yan, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Deep aggregation net for land cover classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 4322

[26] Brian Kurinsky. Power line corridor vegetation management: Clearing a path to reliable electric service using lidar. *Maryville, Missouri, Oct*, 2013. 4322

[27] Chun Liu, Doudou Zeng, Hangbin Wu, Yin Wang, Shoujun Jia, and Liang Xin. Urban land cover classification of high-resolution aerial imagery using a relation-enhanced multiscale convolutional network. *Remote Sensing*, 12(2), 2020. 4322

[28] Leena Matikainen, Matti Lehtomäki, Eero Ahokas, Juha Hyyppä, Mika Karjalainen, Anttoni Jaakkola, Antero Kukko, and Tero Heinonen. Remote sensing methods for power line corridor surveys. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119:10–31, 2016. 4322

[29] Sebastian Pölsterl. Evaluating Survival Models, May 2019. 4326

[30] Sebastian Pölsterl, Nassir Navab, and Amin Katouzian. Fast Training of Support Vector Machines for Survival Analysis. In Annalisa Appice, Pedro Pereira Rodrigues, Vítor Santos Costa, João Gama, Alípio Jorge, and Carlos Soares, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 243–259, Cham, 2015. Springer International Publishing. 4325

[31] M. Shafiqur Rahman, Gareth Ambler, Babak Choodari-Oskooei, and Rumana Z. Omar. Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Medical Research Methodology*, 17(1):60, Dec. 2017. 4326

[32] Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4322

[33] Shuaiang Rong, Lina He, Liang Du, Zuyi Li, and Shiwen Yu. Intelligent detection of vegetation encroachment of power lines with advanced stereovision. *IEEE Transactions on Power Delivery*, 2020. 4323

[34] Xiao-Sheng Si, Wenbin Wang, Chang-Hua Hu, and Dong-Hua Zhou. Remaining useful life estimation–a review on the statistical data driven approaches. *European journal of operational research*, 213(1):1–14, 2011. 4323

[35] Chao Tian, Cong Li, and Jianping Shi. Dense fusion classmate network for land cover classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 4322

[36] J. Wang, C. Li, S. Han, S. Sarkar, and X. Zhou. Predictive maintenance based on event-log analysis: A case study. *IBM Journal of Research and Development*, 61(1):11:121–11:132, 2017. 4323

[37] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019. 4324

[38] D.W. Wanik, J.R. Parent, E.N. Anagnostou, and B.M. Hartman. Using vegetation management and lidar-derived tree height data to improve outage predictions for electric utilities. *Electric Power Systems Research*, 146:236–245, 2017. 4323

[39] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. page 9. 4323, 4325

[40] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 4322, 4323