DINCON 2019 ID: 1582
XIV CONFERÊNCIA BRASILEIRA DE DINÂMICA, CONTROLE E APLICAÇÕES
25 de novembro a 27 de novembro de 2019 – São Carlos/SP

Comparison of Reliability Analysis Regression Methods for Failures in Distribution Systems associated with weather events

Matheus de Souza Sant'Anna Fogliatto[1]
Carlos Dias Maciel
Department of Electrical and Computing Engineering
São Carlos School of Engineering, USP, São Carlos, SP, Brazil
Michel Bessani[2]
Graduate Program in Electrical Engineering, UFMG, Belo Horizonte, MG, Brazil

Abstract. In modern society, if critical infrastructures as the Electrical Power Distribution System do not work as expected, other systems as transports, economics and health care will suffer consequences. The knowledge about the hazards involving that structure is a prominent area of research, and failure models could provide this type of information. However, different techniques can be applied, but not all return proper results and represents the data well. Statistical criteria should be used to determine the best model for the proposed data. In this study, a comparison will be made with distribution families used in a regression model, including weather data as covariates, to determine which has the best fit to construct a failure model for a real Distribution System.

Keywords. Distribution System, Failure Model, Comparison of methods, Reliability Analysis Regression

1 Introduction

Electric Power Distribution System (DS) is responsible for delivering energy to final customers, such as homes, hospitals and industries, and can be seen as a set of nodes and edges. As electricity comes from sources, generally located in substations, any damage on a node or an edge causes a failure event of the DS. Failure events in this type of infrastructure are the interrupt of the energy delivery and can happen because of accidents with animals, vehicles, vegetation or associated with weather events and terrorism, for example.

Failure events are a matter of concern to governments [7]. The statistics provide models and methods that allow the study of the influence of factors in the occurrence of a specific situation, according to the type of event to be analyzed and the data associated with it. An approach is regression methods, which generate a model governed by an equation.

[1]matheusfogliatto@usp.br
[2]mbessani@eee.ufmg.br

Therefore, there are several techniques associated with regression models, and an evaluation of the best alternative for the data to be used is necessary. This study focuses on the comparison of different distribution families for Reliability Analysis Regression (also know as Survival Analysis) [3], on the construction of a model to determine the expectation of a failure or repair occurrence [5], in minutes, associated with weather events [4] covariates.

In Section 2, the data used in this work will be described, as the regression models and the techniques to compare the fitting of each distribution with the data used. Section 3 shows the models and the results of the comparison techniques, followed by a discussion to explain the results shown, determining the choice of the best model for the two types of regression. For last, in Section 4, a conclusion is made, showing the contribution of this work, as the future implementations using the results presented.

2 Materials and Methods

The data used in this work are from a mid-size Brazilian City, with about 550 thousand inhabitants, founded at just under 90 years. The DS presented in Fig. 1 is from this city. Two different datasets were combined to obtain the type of data necessary for the construct of the regression models. First, a dataset of weather data, from the period of 01/01/2012 to 31/12/2014, with daily values for the amount of rain in millimetres, the number of atmospheric discharges and maximum wind speed in kilometre/hour. The second dataset is from failure events and has the date and hour of failure occurrences. From the second data frame, lifetime (time until an occurrence of a failure event) was found in the scale of minutes.

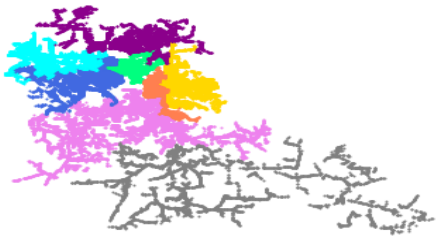


Figure 1: DS composed of eight substations, sixty-five feeders, with 32177 nodes, 32057 edges and a length of 1580 kilometers. Each color represents a substation.

2.1 Data Analysis

2.1.1 Time Data

Fig. 2 present the lifetime of the DS in the form of an histogram. Lifetime is the time until an occurrence of a failure event, or the interval between failures, and was obtained subtracting the date and hour of each failures, using the starting point of the study

(01/01/2012 00:00:00) as the first point (first failure date and hour minus the starting date).

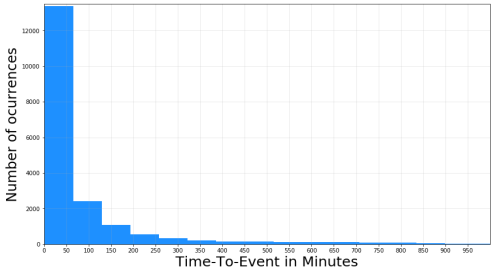
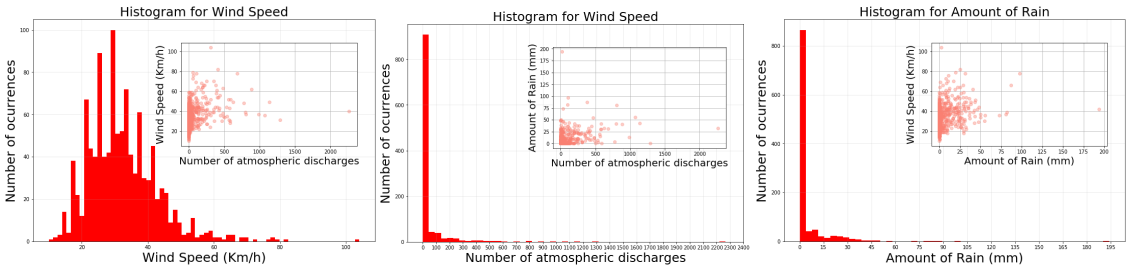


Figure 2: Histogram of time-until-failure ocurrences in minutes. The maximum value registered was 6419 minutes (more than four days) and the minimum value considered was 1 minute, from 18919 registered events. The considered concentration of values for the figure are from 1 to 1000 minutes.

2.1.2 Weather Data

Weather event values are dispersed, as can be seen in Fig. 3, and looks like each weather event is independent. Fig. 4 presents a correlation graphic and reinforces that the weather events are not very related to each other. So, for example, an extreme event for "Amount of Rain" are not necessarily linked to high values of "Atmospheric Discharges".



(a) (b) (c)

Figure 3: Histogram of each weather event and Scatter plot of Weather Event x Weather Event in subplots. The histograms presents a concentration of observations for (b) and (c) at low values. The maximum and minimum value registered for number of atmospheric discharges, wind speed and amount of rain were 2267 and 0, 104 and 10, 194 and 0, respectively.

2.1.3 Statistical Comparison Techniques

For the analysis of the covariates, the p-value was the chosen criteria. Values lower than 0.05 is statistically significant, so, with a covariate return this variable in this margin, it is that it has some impact on the model [1].

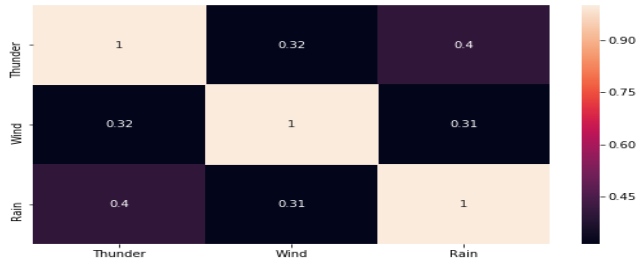


Figure 4: Correlation Matrix of the three weather events used to create regression models. As all values are below 50−80%, it can be considered that each weather event is independent form each other.

For the models, to compares parametric techniques, it can be used non-parametric graphics for Reliability and Cumulative Hazard Functions. The more close a parametric curve is from a non-parametric curve, better was the fitting of the data with the distribution to construct the model. The log-likelihood is used directly in the meaning of comparing models in the presence of uncensored data [2], and lowest values indicate better fitting. The likelihood-ratio test compare the existing model (with all the covariates) to the trivial model of no covariates. Concordance (c-index) is a criterion similar to AUC, and for real data, a value between 0.5 and 0.75 is acceptable. This measure evaluates the accuracy of the ranking of the predicted time [2].

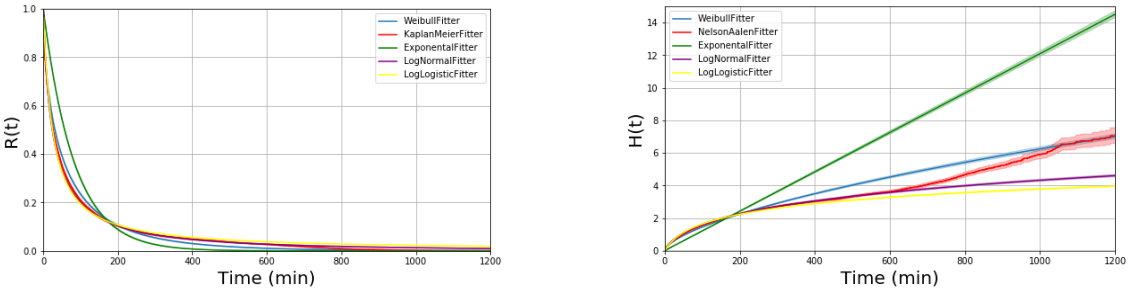
For last, Q−Q (quantile-quantile) plot is a probability plot, using a graphical method for comparing two probability distributions by plotting their quantiles against each other [6]. A good result is if the points in the Q−Q plot approximately lie on the line y = x.

3 Results and Discussion

Survival time for failure events do not present censored data (an incomplete observation), because we have a starting point, that is the first day that a failure happens and we have weather data for that day, and an ending point, the last record of a failure event in the period we have weather data. In the text below, the most common distribution families for parametric techniques [2, 3], Exponential, Weibull, LogNormal and LogLogistic, will be compared. When the covariates (wheater data) are associated, the model is called accelerated failure time (AFT) (For example, Weibull AFT). Some comparison is made without the covariates, but a good result with the regression considering only the survival time generally indicates a good result when incorporating covariates.

The first step on comparison of methods of Reliability Analysis Regression is the plotting Reliability and Cumulative Hazard functions of the desired distribution families with the non-parametric techniques: Kaplan-Meier for $R(t)$ and Nelson-Aalen for $H(t)$. Exponential, Weibull, LogNormal and LogLogistic distributions are plotted with the non-parametric curves, and the results are shown in Fig. 5.

For both curves, Exponential Distribution is discarded because it is not very close to the non-parametric curves. For reliability function, log-logistic and LogNormal curves



(a) R(t) function

(b) H(t) function

Figure 5: (a) Reliability Function graphics comparing the Exponential, LogNormal, LogLogistic and Weibull families distribution with Kaplan-Meier. (b) Cumulative Hazard Function graphics comparing the Exponential, LogNormal, LogLogistic and Weibull families distribution with Nelson-Aalen. For both (a) and (b), the interval of axis x is from 0 to 1200 minutes (most of time-to-event events times are in the interval of 1 to 1000 minutes) to a better visualization.

are superimposed and are closest to the curve generated by the Kaplan-Meier Fitter, but Weibull Fitter curve is tight too. In Cumulative Hazard graph, the three fitters start similar to Nelson-Aalen fitter but begin to diverge at a point. However, the Weibull Fitter returns to the non-parametric curve at the end of the curve.

Weibull, LogNormal and LogLogistic presented an acceptable result in this first analysis. The next step was to compare the criteria's of each model. Tab. 1 shows log-likelihood, concordance and log-likelihood ratio test results.

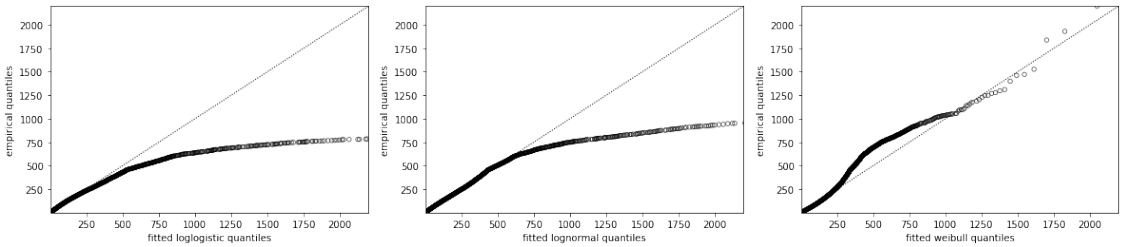
Table 1: Comparison between Weibull, LogNormal and LogLogistic AFT Regression Models

	log-likelihood	Concordance	Log-likelihood ratio test
Weibull	-97230.37360014484	0.56970	1257.66026
LogNormal	-96317.82413640652	0.56970	1333.80861
LogLogistic	-96702.72010305757	0.56960	1519.87994

For the three methods, for concordance index, a very similar value was obtained in the range of 0.55 to 0.75, that is considered the expected value to consider a proper fitting when using real data. Log-Likelihood values are very close to each other, and cannot help to determine each of the three distributions are the most adequate for the desired model. The log-likelihood ratio test has LogLogistic Fitter with a worse result when comparing with Weibull and LogNormal.

The last comparison technique used was QQ-Plot for the three distribution families. In Fig. 6, we can see that only Weibull Fitter presented a great result, with a small deviation of the curve to be followed in an intermediate section and some outliers in the tail of the curve. So, Weibull Distribution is the parametric Reliability Analysis Regression method that best describes our data of weather events associated with the failure of a real

Distribution System.



(a) Empirical versus Fitted quantiles on LogLogistic Distribution. (b) Empirical versus Fitted quantiles on LogNormal Distribution. (c) Empirical versus Fitted quantiles on Weibull Distribution.

Figure 6: QQ-Plot's for LogLogistic(a), LogNormal(b) and Weibull(c) distributions using time-to-fail dataset.

The Weibull Regression parameters for the lifetime of distribution systems associated with the amount of rain (Rain), number of atmospheric discharges (Thunder) and wind speed (Wind) are presented in Tab. 2. The coefficients for each parameter of the model, the standard error (se), the p-value and the 95% confidence interval are the pieces of information presented.

The Weibull distribution has the shape parameter ρ and the scale parameter λ. As proportional hazards are not usual for a failure model, only λ is modelled for the covariates. All coefficients for the weather events have a negative value, that implies that they decrease the lifetime of the system. P-value and the confidence interval shows that all covariates are statically significant for the model, as for all p-values are below 0.05.

Considering "AD" as atmospheric discharges, "W" as wind speed and "R" as the amount of rain, the generated model can be interpreted by Eq. 1 and 2:

$$\lambda(x) = exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right) = e^{(4.94643 - 0.00063 * AD - 0.0241 * W - 0.00346 * R)} \tag{1}$$

$$R(t; x, y) = \exp\left(-\left(\frac{t}{\lambda(x)}\right)^{\rho(y)}\right) = \exp\left(-\left(\frac{t}{\lambda(x)}\right)^{-0.42473}\right) \tag{2}$$

Table 2: Weibull AFT Regression Coefficients

		Coef.	se	p-value	Lower 0.95	Upper 0.95
$\widehat{\lambda}$	Intercept	4.94643	0.03191	<5e-06	4.88389	5.00896
	Thunder	-0.00063	0.00006	<5e-06	-0.00074	-0.00052
	Wind	-0.02410	0.00087	<5e-06	-0.02580	-0.02240
	Rain	-0.00346	0.00078	0.00001	-0.00499	-0.00192
$\widehat{\rho}$	Intercept	-0.42473	0.00536	<5e-06	-0.43524	-0.41422

4 Conclusion

This work proposed a discussion about the relevance and correlation of weather events, which are usually associated as causes for failure events in DSs, and the comparison between Reliability Analysis Regression Methods by graphic evaluations and statistical criterion's.

The weather events amount of rain, wind speed and the number of atmospheric discharges have been proven as statistic significant, and all three have an impact on the occurrence of failures on the analysed Distribution System. Another point addressed was the independence of each other, first by graphics, and after on the significance criteria of the regression method. These results show that create covariates with multiplicative effect (value of two covariates added/multiplied) would not be a right approach.

Another fundamental result presented was that Weibull Distribution is the best model to describe this type of data. For future works, involving, for example, another DSs (more nodes and edges, different city), adding more covariates or make a study about the various causes of failures, we can choose this model knowing that it may fit well with the data.

Acknowledgment

This work was partially support by InSAC: FAPESP: 2014/50851-0, CNPq: 465755/2014-3; COPEL PD 2866-0504/2018 and BPE Fapesp 2018/19150-6.

References

[1] B. Wang et al, The p-value and model specification in statistics, General Psychiatry, (2019), 32:e100081. DOI: 10.1136/gpsych-2019-100081

[2] C. Davidson-Pilon et al, CamDavidsonPilon/lifelines: v0.22.0, (2019), DOI: 10.5281/zenodo.3267531.

[3] E. A. Colosimo and S. R. Giolo, Análise de sobrevivência aplicada, Edgard Blücher, (2006).

[4] F. H. Jufri et al, State-of-the-art review on power grid resilience to extreme weather events: Definitions, frameworks, quantitative assessment methodologies, and enhancement strategies, Applied Energy, 1049-1065, (2019).

[5] M. Bessani et al, A statistical analysis and modeling of repair data from a Brazilian Power Distribution System, 17th International Conference on Harmonics and Quality of Power (ICHQP), (2016).

[6] R. J. Hyndman and Y. Fan, Sample Quantiles in Statistical Packages, American Statistician, vol. 50, no. 4, 361-365, (1996).

[7] S. Brem, Critical infrastructure protection from a national perspective, European Journal of Risk Regulation, vol. 6, no. 2, (2015).