Variable Importance Assessment in Lifespan Models of Insulation Materials: A Comparative Study

Farah Salameh, Antoine Picot, Marie Chabert, Eve Leconte, Anne Ruiz-Gazen, Pascal Maussion

Variable Importance Assessment in Lifespan Models of Insulation Materials: A Comparative Study

F. Salameh[1], A. Picot[1], M. Chabert[2], E. Leconte[3], A. Ruiz-Gazen[3] and P. Maussion[1]

1 : Université de Toulouse ; INPT, UPS ; LAPLACE ; ENSEEIHT, 2 rue Camichel, 31071 Toulouse, France
2 : Université de Toulouse ; INPT, UPS ; IRIT ; ENSEEIHT, 2 rue Camichel, 31071 Toulouse, France
3 : Université de Toulouse ; UT1 ; TSE (GREMAQ), 21 Allée de Brienne, 31042 Toulouse, France

***Abstract* — This paper presents and compares different methods for evaluating the relative importance of variables involved in insulation lifespan models. Parametric and non-parametric models are derived from accelerated aging tests on twisted pairs covered with an insulating varnish under different stress constraints (voltage, frequency and temperature). Parametric models establish a simple stress-lifespan relationship and the variable importance can be evaluated from the estimated parameters. As an alternative approach, non-parametric models explain the stress-lifespan relationship by means of regression trees or random forests (RF) for instance. Regression trees naturally provide a hierarchy between the variables. However, they suffer from a high dependency with respect to the training set. This paper shows that RF provide a more robust model while allowing a quantitative variable importance assessment. Comparisons of the different models are performed on different training and test sets obtained through experiments.**

***Index Terms* — design of experiments, lifespan, modeling, random forest, regression tree, response surface, outliers, twisted pairs, variable importance**

I. INTRODUCTION

THE aerospace industry is moving towards the design of More Electrical Aircrafts (MEA) by replacing heavy mechanical and pneumatic based systems with more electrical based systems [1]-[2]-[3]. This concept offers significant benefits in terms of reliability, much lower operating costs, less impact on the environment, and improved performance [2]. However, the increase in power demand for the electrical equipment supply requires higher voltages and operating frequencies [1], increasing the potential risk of partial discharge (PD) in the insulation systems [4], previously designed for lower voltages. Consequently, the lifespan of electrical insulation materials becomes a key issue for aircraft reliability assessment. In addition to high electrical constraints, other operating stress factors such as temperature, humidity, and mechanical stress contribute to the degradation of the insulating materials [5]. Empirical and physical models have been developed to relate the insulation aging mechanism or lifespan with applied stress factors [6]-[7]-[8]. These models are restrictive since they take into account a single aging factor as in the case of the Arrhenius law, or two factors as in the case of the electrothermal Crine model. In practice, the insulation lifespan is sensitive to numerous factors and to their interactions. Moreover, most of these models include physical parameters related to the studied material, whose estimation requires complex experiments. In recent years, statistical methods have been successfully used in electrical engineering for lifespan modeling based on accelerated aging tests [9]-[10]. These tests consider extreme constraints to speed up the degradation mechanism and to obtain measurable lifespan data [11]. Based on this principle, complete insulation lifespan models are provided in this paper by considering three main aging factors: voltage, frequency and temperature, as well as their interactions. Experiences are organized by Design of Experiments (DoE) [12] and Response Surface (RS) [13] methods. Some extra experiments are also carried out, without constrained levels, for model validation. Then the overall measurements are considered to derive either parametric models based on DoE and RS, or non-parametric models based on recursive partitioning methods as regression trees [14] and random forests (RF) [15]. The common aspect in these different models is the high number of variables (factors and interactions). This paper focuses on methods allowing the assessment of each variable effect and contribution in the resulting lifespan model. This study allows the identification of the least significant variables that can be eliminated, leading to a simpler and more accurate model, with a reduced number of required experiences. The paper is organized as follows: section II describes the experimental setup and the testing methodology. The measured data are analyzed in section III. In sections IV and V insulation lifespan is modeled through parametric and non-parametric methods, with an evaluation of relative errors and variable importance. Finally, conclusions and future works are discussed in section VI.

II. EXPERIMENTAL SETUP AND METHODOLOGY

A. Materials

The tested samples were selected among the most widely used materials in rotating machine wiring insulation for aeronautics applications [9]-[10]. Each sample consists of a twisted pair covered with a double layer of insulating

varnish of Poly-Ether-Imide (PEI) and Poly-Amide-Imide (PAI) with a thermal class of 200°C (Ederfil C200 with a diameter of 0.5mm), as shown in Fig. 1. Twisted pairs are manufactured according to the American National Standard [16].

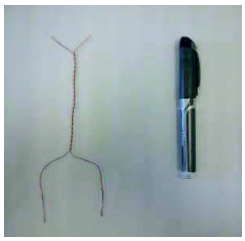

Fig. 1. Twisted pairs EDERFIL C200 as test samples, with a double insulating varnish of Poly-Ether-Imide (PEI) and Poly-Amide-Imide (PAI).

B. Stress factors

The single stress approach offers relative simplicity, but it is inapplicable to real life operating conditions. During its service life, the insulation of rotating machines is subjected to a combination of different stress factors: thermal, electrical, mechanical and environmental [5] - [17], all contributing to reduce its lifetime.

In this paper, a generalized model is presented for insulation lifespan using the multistress approach, with all interactions taken into account. Three main factors are considered: the applied voltage (a periodic square wave with amplitude V), its frequency (F), and the temperature (T). According to [9], the insulation lifespan logarithm (Log(L)) is supposed to follow an inverse power model depending on Log(10V), Log(F) and exp(-bT), with constant b = 4.825×10^{-3} estimated as in [9]. Consequently, these forms are considered in the following lifespan models.

C. Accelerated aging tests

In order to get realistic lifetime measurements, materials are tested under high stress levels, i.e. higher than nominal operation conditions. This procedure, known as Accelerated Life Test, is widely used in aging studies in order to reduce the lifetime of materials under test [11].

This study deals with insulation degradation which is mainly due to PD phenomenon, occurring at high voltages and frequencies. Temperature values vary in a wide range corresponding to the different operating conditions of a rotating machine. They are also chosen within the thermal class of the studied insulation materials. Table I lists the amplitude and frequency ranges of the applied voltage stress, as well as the temperature range.

TABLE I
EXTREME VALUES OF THE THREE STRESS FACTORS

Factors	Minimum Value	Maximum Value
Voltage (kV)	1	3
Frequency (kHz)	5	15
Temperature (°C)	-55	180

D. Test bench

Materials are tested in a climatic chamber where the temperature (T) can be set at the desired value ranging from -55°C to 180°C. A power electronic system generates a square voltage controlled in amplitude (V) and frequency (F). The experimental setup is depicted in Fig. 2.

Thirty-two experiments were carried out. Each one is defined by a combination of the stress values: V, F and T. Eighteen experiments are specified according to a design method that is described in section IV, while the other experiments are carried out with no particular values for V, F and T. Six samples are tested simultaneously for each experiment. Then the failure time of each sample is measured, defining its lifespan at the considered stress levels. The measured lifespans range from 7 s to 1 h 21 mn.

Fig. 2. Climatic chamber and power electronics as a test bench for the insulation materials.

III. STATISTICAL ANALYSIS OF MEASURED LIFESPAN

A. Outlier detection

Outlier detection is a primary step in any modeling task. Outliers are defined as observations whose values deviate from the expected range and may lead to biased modeling results [18]. It is therefore important to identify aberrant lifespan values in the dataset prior modeling. Note that outliers are identified among the six measured lifespans associated to each experiment separately. This can be achieved by using the boxplot graphical tool [18] displayed in Fig. 3. Boxplots give a compact graphical summary about the distribution of a variable, based on a set of order statistics (the median, the first, and the third quartiles). Outliers are observations that fall below the LF (Lower Fence) or above the UF (Upper Fence).

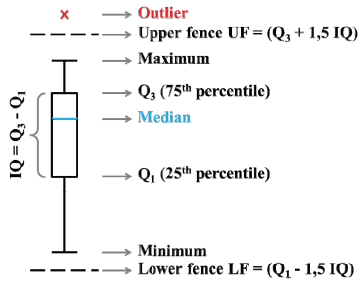

Fig. 3. Boxplot main characteristics.

B. Response form

The measured lifespans are used to derive either parametric or non-parametric models. In parametric models, only a single value is needed to represent the lifespan of each experiment. The mean value can be considered, provided that outliers have been removed. However, the sample median is more robust to extreme values than the sample mean [19]. Therefore, by computing the median of all the repeated measures for each experiment, there is no need for a prior detection of outliers. In non-parametric models, outliers of each experiment are identified and removed. All the remaining lifespans are considered instead of a single value per experiment.

IV. Parametric Models

In this section, the model of the insulation lifespan Log(L) is designed as a linear additive function of the covariates Log(10V), Log(F), exp(-bT) and their interactions. In each studied method, the number of covariates and the required set of experiments composing the training set are specified. The remaining dataset is then used to test the validity of the model. Model parameters are estimated by Ordinary Least Square (OLS) method.

A. Methods

The values of stress factors are specified according to Design of Experiments (DoE) and Response Surface (RS) methods [9]-[10]. For experiment organization purpose, these two methods impose particular levels to each factor. Moreover, these methods consider normalized levels instead of real values.

According to DoE, experiments are organized such that each configuration involves a combination of the levels of the investigated factors [12]. This allows the study of the different effects of the factors simultaneously, increasing accuracy and reducing the number of required experiments. Two levels (±1) are considered in the lifespan DoE model. Consequently, with three factors, $2^3 = 8$ experiments are needed. The lifespan model can be expressed as in (1):

$$Log(L)_{DoE} = M + E_V Log(10V) + E_F Log(F) + E_T exp(-bT) + I_{VF} Log(10V).Log(F) + I_{VT} Log(10V).exp(-bT) + I_{FT} Log(F).exp(-bT) + I_{VFT} Log(10V).Log(F).exp(-bT) \quad (1)$$

RS method [13] is then used to extend the DoE model and to improve its accuracy by adding quadratic forms of the three factors that can also have a significant effect on the response. The lifespan model becomes (2):

$$Log(L)_{RS} = Log(L)_{DoE} + I_{VV} Log(10V)^2 + I_{FF} Log(F)^2 + I_{TT} exp(-2bT) \quad (2)$$

Therefore, three additional levels are required. The design configuration is specified according to Central Composite Design defined by:

- A complete 2^3 DoE design,
- Two axial points situated on the axis of each factor at a distance μ from the design center, defining two extra levels ($\pm \mu$),
- n_0 central points at the design center, i.e. all factors at the 0 level.

n_0 and μ values are set to 4 and $\sqrt{2}$ respectively, so that the obtained design is orthogonal. Thus the total number of required experiments is 18.

B. Required experiments

Table II displays the different configurations of the experiments required by DoE and RS methods. Levels are then defined in Table III.

TABLE II
Levels of the Stress Constraints Required for DoE and RS

Experiences		Level for factor V	Level for factor F	Level for factor T
RS	**DoE**	-1	-1	-1
	DoE	-1	-1	1
	DoE	-1	1	-1
	DoE	-1	1	1
	DoE	1	-1	-1
	DoE	1	-1	1
	DoE	1	1	-1
	DoE	1	1	1
	Axial Point	-√2	0	0
	Axial Point	√2	0	0
	Axial Point	0	-√2	0
	Axial Point	0	√2	0
	Axial Point	0	0	-√2
	Axial Point	0	0	√2
	4 Central Points	0	0	0

TABLE III
Normalized Levels of the Stress Factors

Levels	Log(10V) (kV)	Log(F) (kHz)	Exp(-bT) (°C)
-√2	Log(10*1)	Log(5)	Exp(55b)
-1	Log(10*1.174)	Log(5.872)	Exp(34.82b)
0	Log(10*1.73)	Log(8.7)	Exp(-26.12b)
+1	Log(10*2.554)	Log(12.77)	Exp(-119.74b)
+√2	Log(10*3)	Log(15)	Exp(-180b)

C. Results

Equations (1) and (2) can be seen as linear regression models relating the response vector Y = Log(L) composed of median lifespans with the covariate levels Log(10V), Log(F), etc. composing the covariate matrix X. Let β be the unknown parameter vector to be estimated, thus (1) and (2) can be written in the matrix form: $Y = X\beta$, where β can be estimated by the OLS method.

1) DoE model

The first lifespan model is derived from only 8 experiments according to the DoE method. The model is applied on the remaining 24 experiments composing the test set. Relative errors between predicted and measured responses in the test set range from 0.84% to 234% with an average value of 31%.

The estimated parameters (average lifespan *M*, factor effects, and interaction effects) and the comparison between measured and predicted responses are displayed in Fig. 4. From the bar graph of Fig. 4, it can be observed that voltage and temperature have higher effects than the frequency, which also explains why their interaction is the most influential with respect to the other interactions.

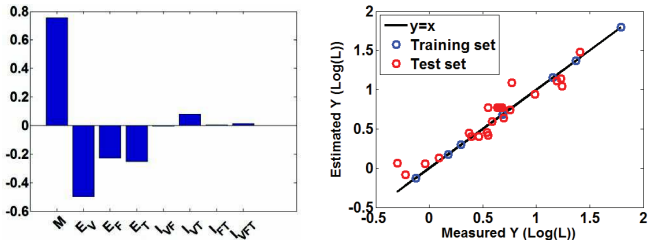

Fig. 4. DoE model: estimation of variable effects (right side) and comparison between measured and estimated lifespans (left side).

2) RS model

The factor effects obtained by DoE model reflect the practical reality, regarding the high influence of voltage and temperature. However, the model seems to be insufficient since some test points present very high errors (>100%). The model is thus extended by adding quadratic terms, leading to RS model. The training set now consists of 18 experiments. The results are depicted in Fig. 5.

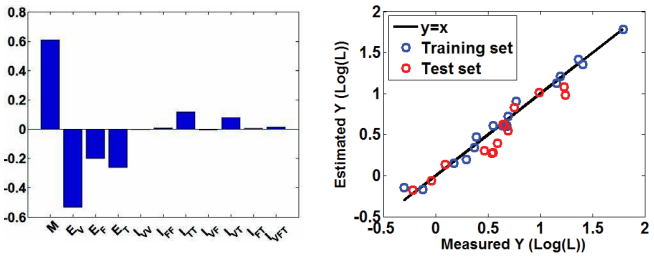

Fig. 5. RS model: estimation of variable effects (right side) and comparison between measured and estimated lifespans (left side).

In addition to the high effects of V, T and their interaction, the RS model reveals a significant contribution of the quadratic term T^2. The maximum and average relative errors computed on the test set decreases down to 53% and 25% respectively.

Therefore, this model is more accurate than the DoE model since it takes more significant effects into account, and it leads to lower errors in the test set.

V. Non-Parametric Models

Previous models assume a linear additive relationship between the response and the predictors. However, it may be of interest to relax these assumptions and to provide a different lifespan-stress relationship with no explicit parametric form. Multivariate non-parametric methods present an alternative approach to linear regression models and are much more appropriate when models include a large number of predictor variables. In the following, non-parametric lifespan models are provided by means of two methods based on recursive partitioning.

A. Regression trees

1) Overview

Classification and regression trees were introduced by Breiman et al. in 1984 [14] for both exploring and modeling categorical (classification) or numeric (regression) data. Trees explain the variation of a single response variable (output) by one or more explanatory variables (inputs). In this study, only regression trees are considered, both predictors and response variables being numeric.

The basic idea behind regression trees is to recursively split the data into smaller and more homogeneous groups. At each node, the splitting explanatory variable and its corresponding threshold value are selected so that the homogeneity of the two resulting groups is maximized. At the end, each leaf is characterized by the mean value of the response variable in the corresponding final group [14]. There are several benefits for using this technique in modeling tasks:

- The relation between the response and the predictor variables is explained through simple if-then rules,
- For a new observation, the response can be easily predicted by following the appropriate path throughout the tree,
- The hierarchical structure of the tree allows to compare the relative importance of the variables,
- Only the most significant predictors are included.

On the other hand, there are two main drawbacks. First, a large number of observations is required so that the algorithm is able to split the data into several groups. Secondly, trees are unstable. Depending on the training set, different trees may be obtained with completely different inputs in the splitting rules, thus leading to completely different interpretations.

2) Application to lifespan modeling

Before applying the regression tree algorithm on the lifespan data, the following rules are defined:

- Inputs: as in RS model, explanatory variables are the main factors (Log(10V), Log(F) and exp(-bT)), their quadratic terms and their interactions. Normalized levels are used.
- Output: the response variable is the measured lifespan logarithm. For each experiment, all repeated

measurements - outliers excluded - are taken into account instead of the unique median value.

- Minimum number of observations per leaf: 15.

Given that regression trees are unstable, prediction is out of the scope of this study. The focus is rather on the modeling part. The tree is computed using all available data (32 experiments). The result is displayed in Fig. 6.

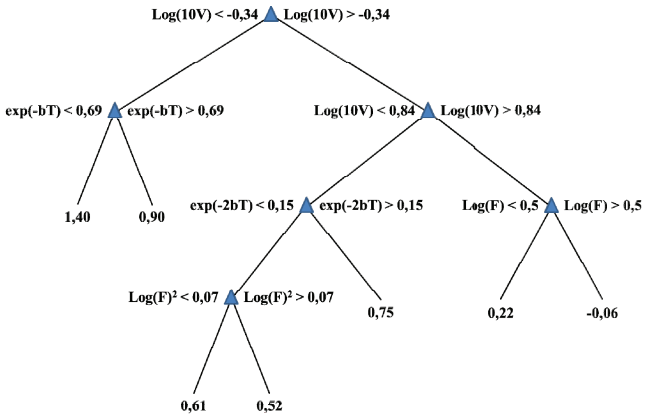

Fig. 6. Regression tree constructed with 32 experiments for insulation lifespan modeling.

3) Discussion

The first analysis of the obtained tree reveals the following observations:

- The voltage is the first splitting variable, meaning that it is the most influent factor,
- The voltage divides the lifespan data into two main subgroups: short lifespans (high voltages, right subtree) and long lifespans (low voltages, left subtree),
- At low voltages, only the temperature has a significant effect on the lifespan,
- Voltage, frequency and temperature appear in the order of their relative importance (V, T then F),
- T^2 is the most influent quadratic term, but is less important than the main factors.

Obviously, the model obtained with the regression tree reveals some similarities with the parametric DoE and RS models: the decreasing effect of V on the lifespan, the relative importance of V, F and T, and the significant effect of the quadratic term T^2 (see Fig. 5).

However, the interaction between V and T does not appear as a significant variable with this tree. On the other hand, this interaction becomes a splitting variable when only the RS training set is used to construct the tree (18 experiments).

Therefore, there is a real dependency between the splitting variables selected by the algorithm and the training set. Conclusions regarding the variable importance are unstable. In order to obtain more robust results, and in an attempt to improve this model, random forests are studied.

B. Random forests

1) Overview

In order to overcome the instability of regression trees and their low prediction performance obtained with a reduced training set, ensemble learning methods were developed. The basic idea is to generate a large number of trees (n_{tree}) and to aggregate their results for more accurate predictions. Based on this principle, random forests (RF) [15] were introduced by Breiman in 2001. Within the past few years, RF have become a very popular and widely-used tool for non-parametric modeling in many scientific domains [20]-[21]. They show high predictive accuracy and are applicable even in high-dimensional problems (where the number of observations n is much lower than the number of predictors p) with highly correlated variables.

In RF, trees are grown similarly to classical regression trees but with two main differences. First, each tree is constructed using a bootstrap sample randomly selected in the sample data. Second, at each node, a given number of input variables (denoted by m_{try}) is randomly chosen and the best split is calculated only within this subset. By default, $m_{try} = p/3$. RF general algorithm is depicted in Fig. 7.

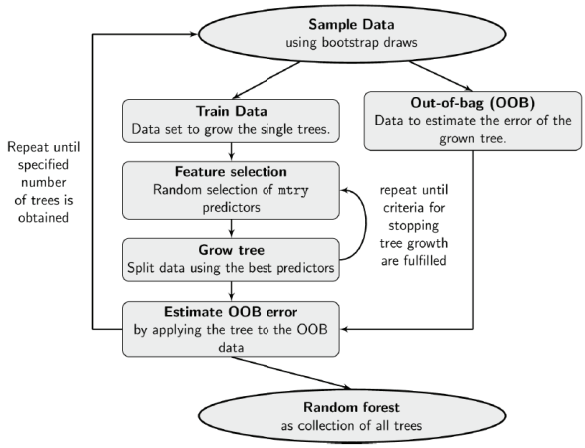

Fig. 7. Random forest general algorithm.

An important feature of RF is the Out-Of-Bag (OOB) sample. An OOB sample is composed by the set of observations that have not been used for building the current tree, and thus can be considered as internal validation data for each tree. OOB samples are used to estimate the RF prediction accuracy and then to quantify the importance of each variable [22]:

- Prediction Mean Squared Error: the accuracy of a random forest prediction can be estimated as in (3):

$$MSE_{OOB} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{\hat{y}}_{iOOB})^2 \qquad (3)$$

where n is the total number of observations, $\bar{\hat{y}}_{iOOB}$ is the average prediction for the i^{th} observation from all trees for which this observation has been OOB.

- Variable importance (VI): the RF algorithm estimates the importance of a variable by averaging, over all the trees, the increase in OOB errors (mean decrease in accuracy) when the observed values of this variable are randomly permuted in the OOB samples, all other variables left unchanged.

2) Variable importance measure in lifespan model

Unlike regression trees, RF are a robust tool for VI assessment. This is demonstrated by examining VI obtained by RF in three different cases. In the following, RF parameters n_{tree} and m_{try} are set to 500 and 3, respectively. As in regression trees, response variable is the measured lifespan logarithm, and the explanatory variables values are the levels of Log(10V), Log(F), etc.

The first RF is generated from all lifespan data (32 experiments). In the second case, only RS experiments are used to generate the RF. VI estimated in these two cases are displayed in Fig. 8 and Fig. 9 respectively. Finally, 50 different RF were generated by randomly selecting a proportion of 2/3 from all the data at each run. For each variable, the computed VI (50 values) are displayed by means of boxplots, Fig. 10.

By comparing the VI magnitudes and medians in the bar diagrams and boxplots respectively, the same conclusions are drawn, meaning that in RF, the measure of VI is robust regardless of the RF training set. It is thus much more convenient to rely on RF rather than regression trees in evaluating variables' relative importance.

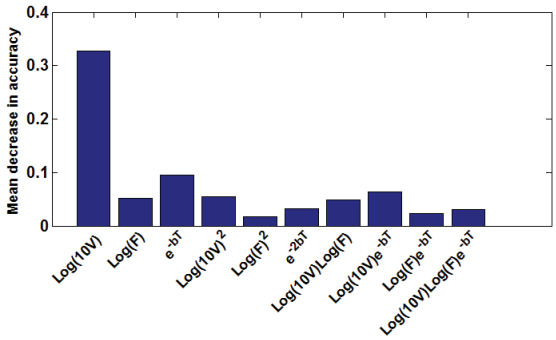

Fig. 8. Variable importance (VI) computed by RF with all the data as a training set.

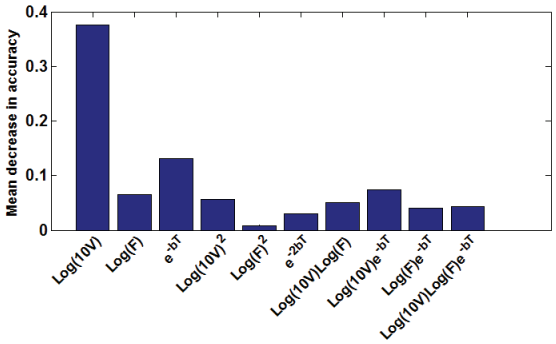

Fig. 9. Variable importance (VI) computed by RF with RS experiments as a training set.

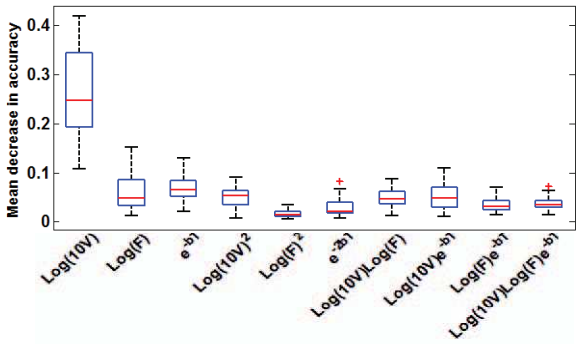

Fig. 10. Variable importance (VI) computed by RF with a randomly selected training set (50 runs).

Once again, voltage and temperature are the most influential factors. On one hand, the interaction between V and T is also the most important with respect to the other interactions. On the other hand, V^2 appears also as an important quadratic term in addition to T^2. This is the only difference with RS variable effects.

3) Error comparison

Table IV summarizes the relative errors computed on the test sets of DoE, RS and RF models (with RF generated from RS training set). Despite all the advantages of non-parametric RF (flexibility, robustness, variable importance quantification), predictions are less accurate than those of the parametric RS model. Note that in all these models, high relative errors correspond to very short lifespans (< 1 min). Fortunately, these points are out of our interest since we are rather concerned in modeling long lifespans.

TABLE IV
TEST POINTS RELATIVE ERRORS

Method	Minimum Error	Maximum Error	Average Error
DoE	0.84%	234%	31%
RS	2.04%	53%	25%
RF - Case 2	3.39%	91%	33%

VI. CONCLUSIONS AND PERSPECTIVES

In this paper, insulation lifespan of twisted pairs covered with varnish is modeled through statistical parametric and non-parametric methods. These different approaches allow the evaluation of the variable importance from different points of view.

In parametric DoE and RS models, the lifespan is expressed as a linear additive function of the predictors and their effects (unknown parameters to be estimated). The most influent factors and interactions are identified as those having the highest estimated effects: the voltage, the temperature, their interaction, and the term T^2.

Although these models are straightforward and accurate, non-parametric regression trees and random forests offer another framework and methodology to model the insulation

lifespan and to rate the variable importance. In regression trees, it is possible to identify the most influent factors by following their hierarchy. However, with different training sets (all lifespan data, and only RS training set), different trees are obtained, leading to different conclusions about the variable importance. The unstable nature of trees is overcome by random forests that combine a large number of trees and average their results. Another advantage of RF is that they allow the quantification of variable importance. The robustness of RF variable importance assessment is demonstrated through three different training sets.

In future works, the RF importance metric (error increase due to variable permutation) will be applied to evaluate the variable importance in DoE and RS models, for a comparison purpose. On the other hand, regression trees and random forests will be used to determine different lifespan models according to the constraint ranges. The results will then be validated by developing a piecewise linear regression model, with the same purpose of obtaining more restricted lifespan models.

REFERENCES

[1] I. Christou, A. Nelms, I. Cotton and M. Husband, "Choice of optimal voltage for more electric aircraft wiring systems", *IET Electr. Syst. Transp.*, vol. 1, no. 1, pp. 24-30, 2011.

[2] L. Fang, I. Cotton, Z.J. Wang and R. Freer, "Insulation Performance Evaluation of High Temperature Wire Candidates for Aerospace Electrical Machine Winding", in Proc. *IEEE Electrical Insulation Conf.*, 2013, pp. 253-256.

[3] I. Christou and I. Cotton, "Methods for Partial Discharge Testing of Aerospace Cables", in Proc. *IEEE International Symposium on Electrical Insulation*, 2010, pp. 1-5.

[4] Y. Jinkyu, B. L. Sang, Y. Jiyoon, L. Sanghoon, O. Yongmin and C. Changho, "A Stator Winding Insulation Condition Monitoring Technique for Inverter-Fed Machines", *IEEE Transactions on Power Electronics*, vol. 22, no. 5, pp. 2026-2033, 2007.

[5] K. A. Sokolija, "A multifactor stress aging model of electrical insulation", in Proc. *Sixth International Conference on Dielectric Materials, Measurements and Applications*, 1992, pp. 374-377.

[6] L. Escobar and W. Meeker, "A review of accelerated test models", *Statistical Science*, vol. 21, no. 4, pp. 552-577, 2006.

[7] G. Mazzanti, "The combination of electro-thermal stress, load cycling and thermal transients and its effects on the life of high voltage ac cables", *IEEE Trans. Dielectr. Electr. Insul.*, vol. 16, no. 4, pp. 1168-1179, 2009.

[8] Z. Li, K. S. Moon, Y. Yao, K. Hansen, K. Watkins, L. Morato and C.P. Wong, "Carbon nanotube/polymer nanocomposites: Sensing the thermal aging conditions of electrical insulation components", *Carbon*, vol. 65, pp. 71-79, 2013.

[9] N. Lahoud, J. Faucher, D. Malec and P. Maussion, "Electrical aging of the insulation of low voltage machines: model definition and test with the design of experiments", *IEEE Trans. Ind. Electron.*, vol. 60, no. 9, pp. 4147-4155, 2013.

[10] A. Picot, D. Malec and P. Maussion, "Improvements on lifespan modeling of the insulation of low voltage machines with response surface and analysis of variance", in Proc. *9th IEEE International Symposium on Diagnostic for Electrical Machines, Power Electronics and Drives*, 2013, pp. 607-614.

[11] J. Pulido, "Using Accelerated Life Testing Techniques for Preventive Maintenance Scheduling", in Proc. *Reliability and Maintainability Symposium*, 2014, pp. 1-6.

[12] R.A. Fisher, *The Design of Experiments*, Edinburgh, U.K.: Oliver and Boyd, 1935.

[13] R.H. Myers and D.C. Montgomery, *Response Surface Methodology*, New York: John Wiley and Sons, 2002.

[14] L. Breiman, J. H. Friedman, R. A. Olshen, and C. G. Stone. *Classification and Regression Trees*, California: Wadsworth, 1984.

[15] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[16] American National Standards Institute, ANSI/NEMA MW 1000-2003, Revision 3, 2007.

[17] P. Nussbaumer, M. A. Vogelsberger, and T. M. Wolbank, "Induction Machine Insulation Health State Monitoring Based on Online Switching Transient Exploitation", *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1835-1845, 2015.

[18] H. Sim, F. F. Gan, and T. C. Chang, "Outlier Labeling with Boxplot Procedures", *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 642-652, 2005.

[19] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection", *WIREs Data Mining and Knowledge Discovery*, vol. 1, pp. 73-79, 2011.

[20] A. Ruiz-Gazen and N. Villa, "Storms prediction: logistic regression vs random forest for unbalanced data", *Case Studies in Business, Industry and Government Statistics*, vol. 1, no. 2, pp. 91-101, 2007.

[21] M. Walschaerts, E. Leconte and P. Besse, "Stable variable selection for right censored data: comparison of methods", 2012, arXiv: 1903-4928.

[22] U. Grömping, "Variable Importance Assessment in Regression: Linear Regression versus Random Forest", *The American Statistician*, vol. 63, no. 4, pp. 308-319, 2009.

AUTHORS' INFORMATION

Farah Salameh got her MSc in electrical engineering from the National Polytechnic Institute of Toulouse (INPT), Toulouse, France, in 2013. She is now a PhD Student at the Laboratory of Plasma and Energy Conversion (LAPLACE) in Toulouse, France. Her work focuses on developing statistical methods for modeling the lifespan of electrical components.

Antoine Picot received the MSc degree in signal, image, speech processing and telecommunications in 2006 and his PhD in automatic control and signal processing in 2009 from the INP Grenoble, France. He is actually an associate professor at the INP Toulouse. He is also a Researcher with the LAPLACE, Toulouse. His research interests are in monitoring and diagnosis of complex systems with signal processing and artificial intelligence techniques.

Marie Chabert received the Eng. degree in Electronics from ENSEEIHT and the M.Sc. degree in Signal Processing from the INPT France, both in September 1994. In December 1997, she received the Ph.D. degree in Signal Processing from the INPT. She is a full Professor in the INPT and in the IRIT (Institut de Recherche en Informatique de Toulouse) laboratory. Her research interests include non-uniform sampling, time-frequency diagnosis, and statistical modeling of remote sensing images.

Eve Leconte received the Eng. degree of Ecole des Mines de Saint-Etienne in 1991, her Master degree in Statistics and Health and her PhD in Biostatistics at University Paris 11 in 1992 and 1995. She is an Associate Professor at Toulouse School of Economics, Université Toulouse 1 Capitole since1996. Her research interests concern the analysis of censored duration data in different settings: survival analysis with multiple events or competing risks, variable selection in survival models, non-parametric estimation in survey sampling.

Anne Ruiz-Gazen got her Master and PhD in Applied Mathematics (Statistics) at Université Paul Sabatier in Toulouse respectively in 1989 and 1993. She is full Professor at Toulouse School of Economics, Université Toulouse 1 Capitole since 2008. Her main research interests are multivariate data analysis and survey sampling.

Pascal Maussion got his MSc and PhD in Electrical Engineering in 1985 and 1990 from INP Toulouse, France. He is full Professor with the University of Toulouse and with the LAPLACE, Toulouse. His research activities deal with control and diagnosis of electrical systems and with the design of experiments for optimization. He is currently Head of Control and Diagnosis group in LAPLACE.