Improving scheduled maintenance by missing data reconstruction: a double-Loop Monte Carlo approach

M. Compare[1,2], F. Di Maio[1,2], E. Zio[1,2,3], F. Carlevaro[4], S. Mattafirri[4]

[1]*Energy Department, Politecnico di Milano, Via Ponzio 34/3, 20133 Milano, Italy*

[2]*Aramis Srl, Via Viviani 8, Milano, Italy*

[3] *Energy Department, Politecnico di Milano, Via Ponzio 34/3, 20133 Milano, Italy*

Chair on Systems Science and the Energetic Challenge, Fondation EDF, Ecole Centrale Paris and Supelec, France

[4] *General Electric- Nuovo Pignone, Via Matteucci 1, Firenze, Italy*

ABSTRACT: *This paper describes a Monte Carlo (MC) based approach for reconstructing missing information in a dataset used by General Electric (GE) for reliability analysis, which contains data coming from field observations at inspection of Gas Turbine components. The approach is based on a combination of Maximum Likelihood Estimation technique to estimate the failure model parameters, Fisher Information Matrix to estimate the confidence intervals on the estimated parameters, and a double-loop MC approach to estimate the missing equivalent starts (i.e., data of turbine state without the relative equivalent starts). The proposed methodology reduces the uncertainty in the estimation of the parameters of the turbine. Results of the application of the novel approach to a real industrial dataset are discussed along with a sensitivity analysis for the quantification of the robustness of the methodology to deal with different sizes of datasets.*

1 INTRODUCTION

Gas Turbines (GTs) are complex and expensive industrial systems, which generate revenue when they are in working state, only. From this, it emerges the relevance for GT operation of optimizing maintenance according to a number of possibly conflicting criteria, the minimization of the system unavailability being surely among them.

The economic importance of maintenance is one of the main justifications of the dramatic increase in the outsourcing of equipment maintenance experienced in recent years, especially in case of complex and expensive products such as the GTs (Murthy & Asgharizadeh, 1999). Other reasons are the fast technological advancements and the more stringent safety and environmental legislation in production processes, which call for increasing needs of well-trained maintenance workforce, and a more specific maintenance management capability. This also increases the cost of maintenance, and makes small maintenance departments within the owner's organization not economically competitive with maintenance outsourcing (i.e. the scale of operation is too small) (Wang, 2010).

Moreover, the relative affordability of on-line monitoring technology has led to a growing interest in new maintenance paradigms such as the CBM and PrM (Zio & Compare, 2013): the possibility of monitoring the system and, by so doing, collecting information on its conditions is used to both identify problems at an early stage and predict their evolution in the future. This gives the possibility of building an agile logistic network, which allows for a significant reduction in the system unavailability by providing the right part in the right place at the right time, with the necessary resources to perform the maintenance (Pipe, 2008). This has boosted the interest in new maintenance service contracting models (i.e., vendor logistics support, integrated operational support, service by the hour or power by the hour contracts, depending on the industrial field) in which the owner (i.e., who buys GTs from GE to generate revenue) directly buys the availability of the system, rather than the system itself (Hess et al. 2005). In business terms, this yields to the owner the advantages of *i)* transferring the risk of unavailability (e.g., business interruption (Amato et al. 2011)) to the manufacturer who pays penalties when the performance under contract is not guaranteed, *ii)* streamlining its own maintenance department, and *iii)* simplifying the business model, in where there will be a single cost item for the asset management (Pipe, 2008).

This situation, i.e., need for maintenance outsourcing and power by the hour contracting, has led to an increasing interest by the owner in the maintenance service contracts offered by the GT manufacturers (i.e., GE), first because it may be the mandatory condition to sell the GTs, and then because it may be an important source of

income, with new opportunities arising if GE is able to sell added values by taking over parts of clients' business risks and other (financial) burdens. In this way, a contractor can diversify his ''product'' range and may be able to achieve a higher profit.

From the considerations above, it clearly emerges that a reduction of the epistemic uncertainty in the GT behavior is mandatory to have a profitable service contract, as this allows reducing the risk associated to the contract and, thus, defining profitable terms and conditions for purchasing the equipment and associated maintenance.

In this context, GE has started a data collection project to populate a Part Life DataBase (PLDB) containing quantitative and qualitative data on a number of components and degradation mechanisms coming from field observations at inspection of GTs produced by GE. These data are**, then,** used to infer the future component failure behavior and identify the corresponding optimal maintenance decision that minimizes a cost objective function. Moreover, PLDB can be used to produce the cost estimate of maintenance service contracts to be signed with customers.

The data mainly come from the GT operated by GE within maintenance service contracts, for which:
I. The customer sends GE the component removed from the system upon a maintenance inspection. GE experts check the component and qualitatively classify its health state into 'Good', and 'Heavy' (binary failure model).
II. The component ID, the information about its working time and the result of the inspection are stored in a PLDB for reliability life data analysis.
III. Components that can be fixed (with different repair procedures according to the entity of the damage) are sent back to the custumer in an As Good As New (AGAN) state.

However, there are additional inspection data coming from GTs produced by GE, but not directly operated by GE. This entails that GE does not precisely know the operating conditions of these turbines, and, thus, information relevant to predict the failure time is missing. For example, the Low Cycle Fatigue (LCF) crack propagation mechanism affecting the GT nozzles heavily depends on the equivalent starts, rather than on the operating time, which are not traced by the GT owner. **It is worth mentioning that the equivalent starts are calculated by resorting to a function that, for confidentiality, is not reported here and maps the number of normal, anomalous, fast and slow start-ups experienced by the turbine by differently weighting them into an equivalent value of starts.** Thus, when GE experts perform maintenance, they can only store the degradation state of the nozzles, with missing equivalent starts. This additional dataset represents a useful source of information, which GE would add on the PLDB to not to discard it.

A number of approaches to the missing data reconstruction problem have been proposed in the literature. These can be divided into two classes:
- The first class includes the approaches consisting in developing models based on Artificial Neural Networks (ANN) (Hines et al., 1996), Principal Component Analysis (PCA) (Luh & Lin, 2011), Auto-Associative Kernel Regression Methods (AAKR) (Baraldi et al., 2012), Fuzzy Similarity methods (Baraldi et al., 2013) and similarity measures computed taking the correlation (Kim et al., 2005) and entropy (Brock et al., 2008) of information.
- The second class encompasses the approaches that infer the missing data by statistical values, e.g., the mean value of the available data (Schafer, 2002).

However, the performances of these approaches tend to decrease when applied to datasets containing few complete data, especially in the latter case.

In this context, a novel procedure to exploit the statistical evidence gathered by GE has been proposed in (Di Maio et al., 2014), which also allows improving the reliability predictions. In details, the approach proposed in (Di Maio et al., 2014) is based on a combination of i) Maximum Likelihood Estimation (MLE) technique to estimate the failure model parameters of the turbine components based on the information contained in a complete PLBD (i.e., pairs of components status and equivalent starts), ii) the Fisher Information Matrix to estimate the confidence intervals of the estimated parameters and iii) on a double-loop Monte Carlo (MC) approach to propagate these uncertainties and estimate the missing equivalent starts of an incomplete dataset (i.e., records of turbine status without the relative equivalent starts).

The main benefit is the integration of the information contained in the complete and reconstructed PLDB to reduce the uncertainty in the estimation of the parameters of the GE turbine binary degradation model, and, thus, to support the business department in making more informed decisions.

This approach is first recalled with additional theoretical details. The illustration is based on a case study derived from the GE data, which have all been re-scaled, and a few have even been randomly changed to protect

the intellectual property of GE. These data are different from those used in **(Di Maio et al., 2014). Then, t**he approach is discussed along with a sensitivity analysis for the quantification of the robustness of the methodology to deal with different sizes of PLDBs.

The reminder of the paper is as follows. Section 2 presents the case study; Section 3 illustrates the basics of the MLE technique to estimate the parameters and their uncertainties of the GE turbine binary degradation model; Section 4 presents the novel approach based on a double-loop MC to propagate the parameters uncertainties and estimate the missing equivalent starts of the incomplete dataset. Finally, Sections 5 and 6 show the reduction of the uncertainty of the GT degradation model parameters that is allowed by the acquired statistical evidence gathered by reconstructing the missing equivalent starts and a sensitivity analysis for the quantification of the robustness of the methodology to deal with different sizes of PLDBs, respectively.

2 CASE STUDY

The available complete dataset, i.e., the PLDB, consists of a set N=30 records regarding field observations of the nozzle system of turbines operated by GE. Each record contains (see Table 1):
1. The machine (i.e., turbine) ID;
2. The inspection time (i.e., equivalent starts);
3. The state of the component ('Working' or 'Failed').
4. The time at which the observation started, to give due account to the left censored data;

When the i-th machine with n=22 nozzles is inspected, a number n_s^i and n_f^i of nozzles can be found in 'working' and 'failed' states, respectively. In the latter case, records are referred to as interval censored records, whereas in the former case, records are right censored.

Beside the N=30 complete data, another dataset is available containing the data of M=32 additional machine inspections with unknown equivalent starts, i.e., the degradation state of every nozzle in the turbine is known, but the inspection time at which this situation has been encountered is missing.

In this context, the goal is to:
- Develop a method based on the Maximum Likelihood Estimation (MLE) technique to estimate the parameters of the binary degradation model, based on quantitative and qualitative data coming from field observations at inspection;
- Implement a technique based on the Fisher Information Matrix to estimate the confidence intervals on the estimated parameters;
- Develop and implement a technique for propagating the estimated uncertainties (confidence intervals) onto the probabilities of occupying the degradation states over time;
- Develop a technique to estimate the missing equivalent starts to be associated to the recorded inspection outcomes.

Notice that the data considered in this paper and the corresponding results shown are different from the real ones from GE field experience, in respect of GE policy on confidential information.

Starting time of operation	Inspection time	Number of safe components	Number of failed components
	32.7 [units of time]	12	10
	missing	8	14
	…	…	…

Table 1: example of data in PLDB

3 ESTIMATION OF THE MODEL PARAMETERS

The aim of the first task of the work is to exploit the evidence contained in the available dataset with N=30 complete data to estimate the values of the parameters of the stochastic model describing the degradation mechanism under study. The assumption is that the transition from 'working' to 'failed' state obeys a two-parameter $\underline{\theta} = \{\beta, \eta\}$ Weibull distribution, where β is the shape factor and η the scale factor. This latter represents the number of equivalent starts before which 63% of turbine nozzles fail.

Resorting to the Maximum Likelihood Estimation (MLE) technique (Kendall & Stuart 1979), the LogLikelihood function L can be written as:

$$L(\beta,\eta) \propto \log\left(\prod_{i=1}^{N} F(t_i|\beta,\eta)^{n_f^i} R(t_i|\beta,\eta)^{n_s^i}\right) \qquad (1)$$

where $R(t|\beta,\eta) = e^{-\left(\frac{t}{\eta}\right)^{\beta}}$ is the reliability of the n_s^i nozzles of the i-th machine which is found in 'working' state

upon inspection, whereas $F(t|\beta,\eta) = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}}$ is the unreliability of the n_f^i nozzles that are found in 'fault' state. In this respect, notice that the available dataset contains right-censored data in case of survived components, and interval censored data (where the intervals in which the failures occurred start at time t=0) in case of failed components (Zio, 2007).
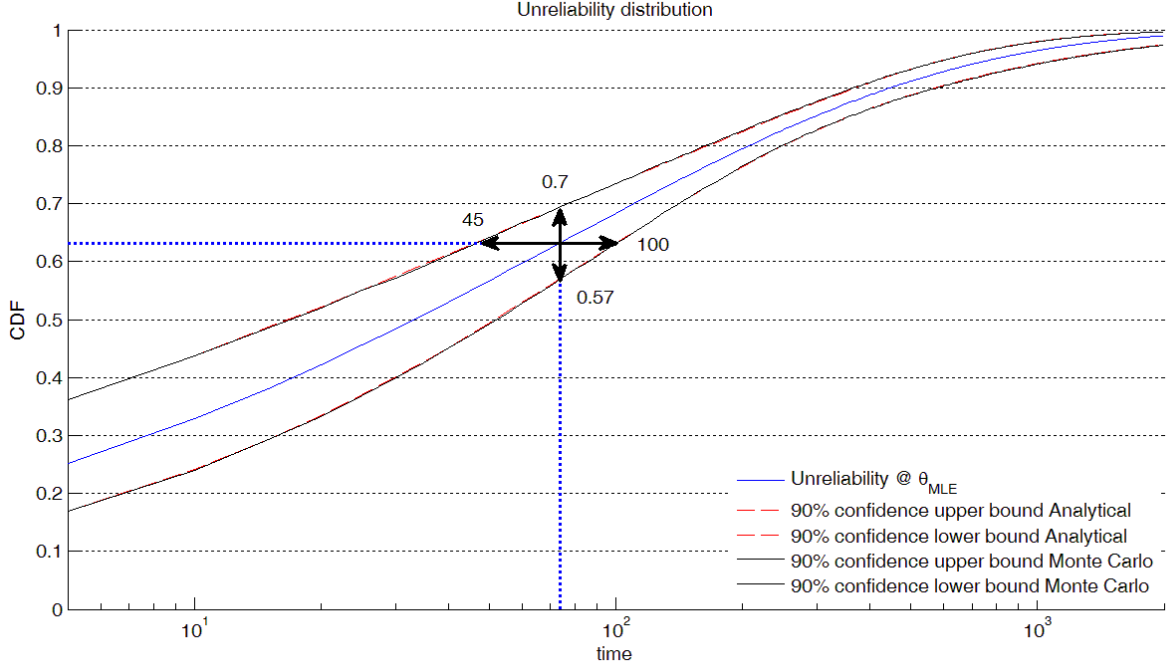

Figure 1: System unreliability distribution with corresponding lower and
upper bound due to the uncertainty in the Weibull parameters.

The application of the MLE technique to the PLDB data relevant to the considered degradation mechanism yields the parameter values $\hat{\underline{\theta}} = \{\hat{\beta},\hat{\eta}\} = \{0.46, 73.67\}$. Figure 1 shows the corresponding Cumulative distribution Function (cdf). Notice that the data which this Figure refers to are derived from the real GE data, which have all been re-scaled, and a few have even been randomly changed to protect the intellectual property of GE.

3.1 *Uncertainty of the model parameters*

An estimation of the confidence intervals on the parameters of a stochastic 2-parameter model can be obtained by exploiting the asymptotic properties of the ML estimator $\hat{\underline{\theta}} = \{\hat{\beta},\hat{\eta},\}$. In fact, if $\hat{\underline{\theta}}$ is a unique solution to the maximization problem of $L(\beta,\eta)$, then (Papoulis & Pillai 2002, Kendall & Stuart 1979):

i) $E(\hat{\theta}) \to \theta$ \qquad\qquad (2)

ii) $(\hat{\sigma}_{\beta},\hat{\sigma}_{\eta}) \to (\sigma_{\beta},\sigma_{\eta})_{CR}$ \qquad\qquad (3)

where $(\sigma_{\beta},\sigma_{\eta})_{CR}$ are the Cramer-Rao lower bounds (Papoulis & Pillai 2002) of the standard deviations of any unbiased estimator of $\underline{\theta} = \{\beta,\eta\}$. That is, any choice of an unbiased estimator of $\underline{\theta}$ brings a variability in the ML estimate which is necessarily larger than $(\sigma_{\beta},\sigma_{\eta})_{CR}$.

iii) $\hat{\underline{\theta}}$ is asymptotically normal, that is, $\hat{\underline{\theta}} - \underline{\theta} \approx N(0, J^{-1}(\underline{\theta}))$. (4)

Where $J(\underline{\theta})^{-1}$ is the inverse of the expected Fisher information matrix $J(\underline{\theta})$ whose generic (i,j) entry, $i, j = 1,2$ is (Kendall & Stuart 1979):

$$J_{ij} = E\left(\frac{\partial \log L(\underline{\theta})}{\partial \theta_i} \cdot \frac{\partial \log L(\underline{\theta})}{\partial \theta_j} \right)$$ (5)

where $L(\underline{\theta}) = L(\beta, \eta)$ is the likelihood function, and $\theta_1 = \beta$ and $\theta_2 = \eta$.

In the reference case study, the estimated standard deviations $\hat{\underline{\sigma}} = \left(\hat{\sigma}_\beta, \hat{\sigma}_\eta \right)$ of $\hat{\underline{\theta}} = \left\{ \hat{\beta}, \hat{\eta}, \right\}$ are given by the entries (1,1) and (2,2) of the matrix $J(\underline{\theta})^{-1}$, respectively:

$$\begin{bmatrix} E\left(\left[\dfrac{\partial \log L(\beta,\eta)}{\partial \beta} \right]^2 \right) & E\left(\dfrac{\partial \log L(\beta,\eta)}{\partial \beta} \cdot \dfrac{\partial \log L(\beta,\eta)}{\partial \eta} \right) \\ E\left(\dfrac{\partial \log L(\beta,\eta)}{\partial \eta} \cdot \dfrac{\partial \log L(\beta,\eta)}{\partial \beta} \right) & E\left(\left[\dfrac{\partial \log L(\beta,\eta)}{\partial \eta} \right]^2 \right) \end{bmatrix}^{-1}$$ (6)

It is worth noticing that under mild regularity conditions, the matrix in Eq. (6) is equal to:

$$\begin{bmatrix} -E\left(\dfrac{\partial^2 \log L(\beta,\eta)}{\partial \beta^2} \right) & -E\left(\dfrac{\partial^2 \log L(\beta,\eta)}{\partial \beta \partial \eta} \right) \\ -E\left(\dfrac{\partial^2 \log L(\beta,\eta)}{\partial \beta \partial \eta} \right) & -E\left(\dfrac{\partial^2 \log L(\beta,\eta)}{\partial \eta^2} \right) \end{bmatrix}^{-1}$$ (7)

This allows us giving a rough, informal interpretation of Fisher information. In fact, notice that in general:

$$\frac{\partial^2 \log L(\underline{\theta})}{\partial \theta_i^2} = \frac{L'(\underline{\theta})}{L(\underline{\theta})}$$ (8)

Then, the second derivatives give a measure of how quickly the log-likelihood changes when we change the parameter values in the proximity of the ML estimate: then, if the Fisher information is large, the distribution changes quickly when we move far from the MLE value of the parameter. This entails that we should be able to properly estimate $\hat{\underline{\theta}}$ based on the data. On the other hand, if Fisher information is small, this means that the distribution is 'very similar' to distributions with parameter not so close to $\hat{\underline{\theta}}$ and, thus, more difficult to distinguish; then, our estimation will be worse.

On the other side, for a positive definite parameter (e.g., the parameters of the Weibull distribution), the normal approximation is sometimes unsatisfactory because the distribution centred in the MLE values may have the tails in the negative parts of the hyper-space, thus giving negative lower bounds of the confidence interval. In this case, the delta method (Meeker &Escobar, 1995) is adopted to improve the quality of the approximation by applying a suitable re-parameterization of $\hat{\underline{\theta}}$ (i.e., a transformation of the parameters to a new scale). In details, the delta method states that we can work, instead of $\underline{\theta}$, with a differentiable non-zero valued function $\phi(\underline{\theta})$, which is again asymptotically normally distributed around $\phi(\hat{\underline{\theta}})$, with adjusted co-variance matrix $\hat{\Sigma} = \phi'(\hat{\underline{\theta}})^T J^{-1}(\hat{\underline{\theta}}) \phi'(\hat{\underline{\theta}})$, where $\phi'(\hat{\underline{\theta}})$ is the diagonal matrix of the partial derivatives of ϕ with respect to the parameters in $\underline{\theta}$, calculated in $\hat{\underline{\theta}}$. That is, $\phi(\hat{\underline{\theta}}) \approx N(\phi(\hat{\underline{\theta}}), \hat{\Sigma})$. Following (Compare et al., 2014), in this work we consider the function $\phi(\theta_i) = \ln(\theta_i)$ for those parameters with negative tails, otherwise $\phi(\theta_i) = \theta_i$.

3.2 *Confidence intervals on the probability plot*

In this case study, the values of $\hat{\underline{\sigma}} = \left(\hat{\sigma}_\beta, \hat{\sigma}_\eta \right)$ are 0.0650 and 16.34, respectively. These have been estimated by applying the bootstrap-based procedure proposed in (Spall, 2005). Figure 2 shows the bi-variate normal distribution of the MLE $\hat{\underline{\theta}}$.

The uncertainty in the MLE values must be propagated through the Weibull stochastic model. To this aim, we have applied the following MC sampling procedure (Zio, 2013):

i) Partition the time axis into C small channels of length Δt.

ii) Sample (β, η) from $\phi(\hat{\theta}) \approx N(\phi(\hat{\theta}), \hat{\Sigma})$;

iii) For each channel i of the time axis, compute $R(t_i | \beta, \eta)$, where $t_i = i \cdot \Delta t$, $i=1,\ldots, C$.

iv) Repeat steps ii)-iv) a large number of times (e.g., 500 in this work).

v) For each time channel i, find the 5th and 95th percentiles of the gathered statistics, i.e., the pointwise double-sided 90% confidence interval.

Figure 1 shows the CDF of the unreliability and its pointwise double-sided 90% confidence interval. The width W_η of the double-sided 90% confidence intervals for η is equal to $W_\eta = 100 - 45 = 55$, whereas the width W_U of the double-sided 90% confidence intervals for the unreliability at $t = \hat{\eta}$ is equal to $W_U = 0.7 - 0.57 = 0.13$. These widths are also reported in Figure 1.

Finally, the estimations of the pointwise double-sided 90% confidence intervals provided by the MC procedure described above are also cross-checked against the analytical results provided by (Nelson, 1982): Figure 1 shows the solutions of the two approaches being very close to each other.
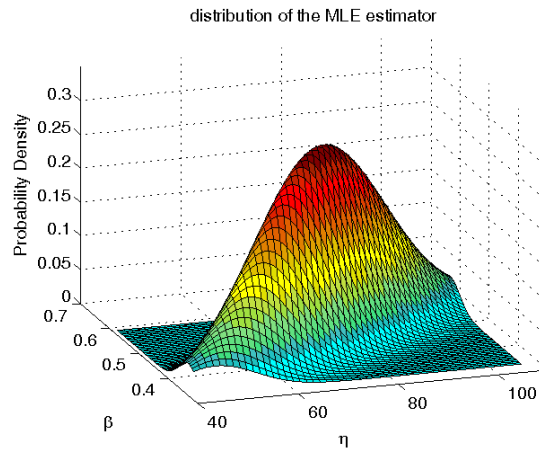

Figure 2: Distributions of the MLE estimator.

4 ESTIMATION OF THE EQUIVALENT STARTS

The aim of this Section is to propose a procedure to infer such equivalent starts, on the basis of the statistical evidence made up of the N complete data. In our approach, we make use of a MC based approach for reconstructing missing information in the PLDB: as anticipated above, a number M of additional turbine inspection outcomes have been collected by GE, which report the degradation state at inspection of each of the $n = n_f^i + n_s^i$ components of the i-th machine nozzles, but not the time (expressed in equivalent turbine starts) at which that inspection has been performed.

From the statistical standpoint, every machine inspection can be regarded as the result of n repeated trials, each one consisting in testing the health state of the nozzle after it has been working for a time interval $[0, t]$. For each trial, two mutually exclusive events can occur: 'nozzle is in failed state', and 'nozzle is in working state'. For the sake of simplicity, we assume that the n trials are independent on each other. From the practical point of view, this means that the evolution of the degradation mechanism affecting any nozzle does not influence the evolution of any other degradation process relevant to any other nozzle of the turbine.

The probability of having a failed nozzle in a single trial reads:

$$p = F(t|\beta,\eta) = 1 - e^{-\left(\frac{t}{\eta}\right)^\beta} \qquad (9)$$

Simple analytical manipulations of Eq. (9) yield:

$$t = \eta \sqrt[\beta]{-\log(1-p)} \qquad (10)$$

Eq. (10) allows us relating t with the degradation states recorded on the i-th machine at time t. Data in the original database allows us to calculate the ML estimators $\hat{\underline{\theta}} = \{\hat{\beta}, \hat{\eta},\}$ of $\underline{\theta} = \{\beta, \eta,\}$: these are normally distributed random variables, as shown in Section 3. It is important to notice that also p is a random variable, as it depends on the inspection outcome n_f^i of the i-th machine at time t, as shown in the following. Therefore, t turns out to be a random variable, that is a function of the random variables β, η and p.

Our objective is to find the ML estimate $\hat{p}^i$ of the probability p^i of having at inspection of the i-th turbine, $i=1,\ldots,N$, a given value n_f^i of failed components and, thus, $n_s^i = n - n_f^i$ of safe components.

To do this, first notice that the number of failed components N_f^i is a binomially distributed random variable. Therefore,

$$P(N_f^i = n_f^i) = \binom{n}{n_f^i} p^{i \, n_f^i} (1-p^i)^{n-n_f^i} \qquad (11)$$

Then, given the outcome n_f^i of N_f^i, the ML estimate of the probability p^i is given by the value $\hat{p}^i$ that maximizes the likelihood function $L(n_f^i \mid p)$:

$$L(n_f^i \mid p) = \binom{n}{n_f^i} p^{i \, n_f^i} (1-p^i)^{n-n_f^i} \qquad (12)$$

from which we get that the maximum likelihood estimate $\hat{p}^i = \dfrac{n_f^i}{n}$ (Papoulis & Pillai, 2002).

The second step of the proposed procedure aims at describing the uncertainty in the ML estimator $\hat{p}^i$, which is a random variable with its own distribution. In this respect, notice that the probability of having a number N_f^i of observations of 'failed' components not larger than n_f^i is:

$$P(0 \le N_f^i \le n_f^i) = \sum_{k=0}^{n_f^i} \binom{n}{k} p^{i \, k} (1-p^i)^{n-k} \qquad (13)$$

with $0 \le n_f^i \le n$. Being p a random variable, the probability that $\hat{p}^i \ge p_u^i$ is given by (Birolini, 2004):

$$P\left(\hat{p}^i \ge p_u^i\right) = \sum_{k=0}^{n_f^i} \binom{n}{k} p_u^{i \, k} (1-p_u^i)^{n-k} \qquad (14)$$

Thus, we can assert that $\hat{p}^i$ is distributed as an incomplete Beta function $I_{1-p}\left(n-n_f^i, n_f^i +1\right)$ (Kendall & Stuart 1979) of Eq. (14). We can, now, sample a possible realization of p for a given n_f^i from this distribution and, any outcome n_f^i we can exploit Eq. (10) to estimate its missing equivalent inspection time t, by sampling α, β and p from the respective distributions.

The following double-loop MC sampling procedure is implemented: the inner loop finds a possible realization of the missing equivalent starts t for an inspection on the i-th turbine with n_f^i failed nozzles by:

i) **Sampling (β, η) from $\phi(\hat{\theta}) \approx N(\phi(\hat{\theta}), \hat{\Sigma})$;**

ii) **Sampling p from $I_{1-p}\left(n-n_f^i, n_f^i +1\right)$;**

iii) **Computing t with Eq. (10).**

The outer loop repeats steps i) to iii) a large number of times (i.e., 10^4) to buildg the probability distribution of t and to calculate its percentiles (e.g., 50th percentile), to be assumed as reconstructed missing equivalent starts.

For example, Figure 3 shows the CDF of the missing time corresponding to the inspection outcome $n_f^i = 16$ and $n=22$.

5 ESTIMATION OF THE PARAMETERS ON THE MERGED DATASET

Hereafter, the merged dataset containing $N+M$ turbine inspections is considered, where the equivalent starts t corresponding to the M turbine inspections are taken from the corresponding distribution estimated as in Section 4. In this case, the Log-Likelihood function L can be written as:

$$L(\beta,\eta) \propto \log\left(\prod_{i=1}^{N+M} F(t_i|\beta,\eta)^{n_f^i} R(t_i|\beta,\eta)^{n_s^i} \right) \qquad (16)$$

where $R(t|\beta,\eta) = e^{-\left(\frac{t}{\eta}\right)^{\beta}}$ is the reliability of the n_s^i nozzles of the i-th machine which are found in 'working' state upon inspection, whereas $F(t|\beta,\eta) = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}}$ is the unreliability of the n_f^i nozzles that are found in 'fault' state. If $i \leq N$, then t_i is the equivalent starts value reported in the original complete dataset, otherwise, t_i is picked from the distribution of t corresponding to n_f^i.
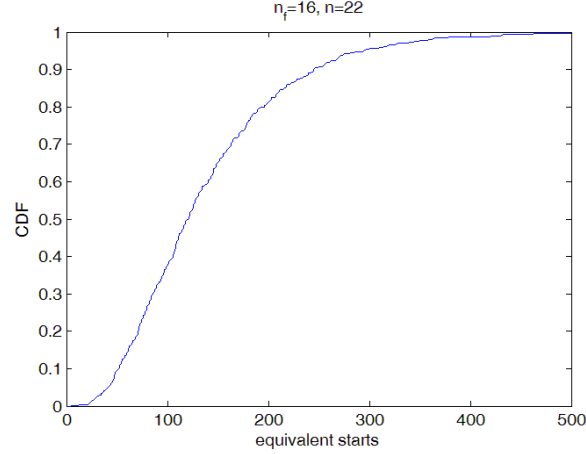

Figure 3: CDF of the missing time corresponding to the inspection outcome n_f=16 and n=22.

In turn, we can randomly sample the M missing times from the corresponding distributions and ML-estimate, through the procedure detailed in Section 3), the Weibull distribution parameters on the merged original and reconstructed datasets of $N+M$ machines. If we repeat a large number of times (e.g., 10^4) such procedure of randomly select the M missing times and collect the ML-estimate of the reconstructed datasets of $N+M$ machine, then we can investigate the variability of the new ML estimator. In fact, every dot in Figure 4 represents a MLE corresponding to a dataset of $N+M$ machine inspections, in which the M missing times are MC sampled.
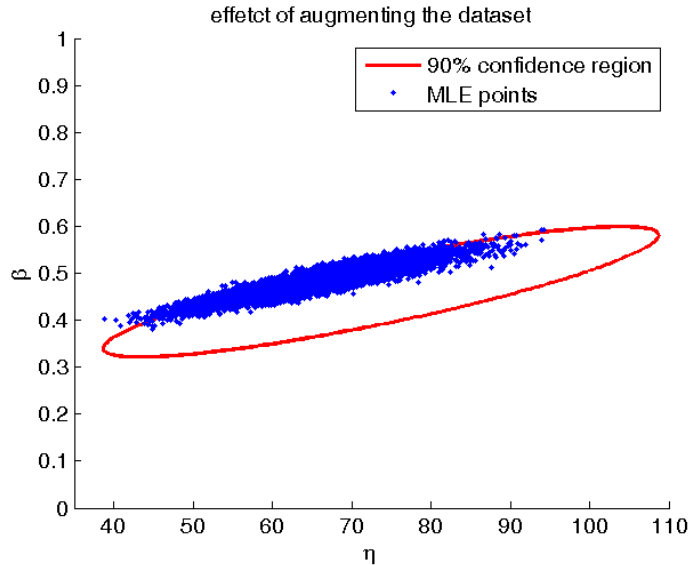

Figure 4: MLE of the Weibull distribution parameters from the merged original and reconstructed datasets of $N+M$ machine inspections, in correspondence of 10^4 sampled M-ple of missing times. The 10^4 ML-estimates are compared to the 90% confidence region relevant to the original N-machine inspection dataset.

From Figure 4, it clearly emerges that the uncertainty in the ML-estimator related to the augmented dataset is smaller than that in the estimator related to the original dataset. In fact, whatever the M missing times are, as long as these are coherent with the corresponding inspection outcomes, the MLE parameters are within the area identified by the blue dots, which is smaller than the 90% confidence area (Wald area, Meeker & Escobar, 1995) of the MLE parameters inferred from the original dataset of N machine inspections. From the collected MLEs, we can also calculate the mean value of the new ML estimator $\hat{\theta} = (0.4833, 65.5286)$.

An additional advantage of the proposed procedure lies in that the strategy to assign the missing times can be defined to accommodate the risk proneness of the decision maker (DM). For example, if the DM needs being conservative (risk averse), then every machine with missing time can be given the first quartile (i.e., 25th percentile) of the distribution corresponding to its inspection outcome, whereas the third quartile (i.e., 75th percentile) can be chosen on the opposite case.

Figure 5 shows the distributions of the ML estimators in correspondence to the three quartiles of the missing data distributions. If we compare the distributions along the first two columns of Figure 5, we can notice that there is a slight change in the shape parameter and a more pronounced shift of the distributions of the scale factor towards lager values of equivalent starts.

Figure 5 also shows that whichever the quartile is, the variance of the estimators inferred from the merged dataset is smaller than that of those inferred from the original dataset (PLDB of N machine inspections). This yields a reduction in the width of the pointwise double-sided 90% confidence intervals for η and for the unreliability at $t = \hat{\eta}$ (plots in the last column in Figure 5, compared to Figure 1)
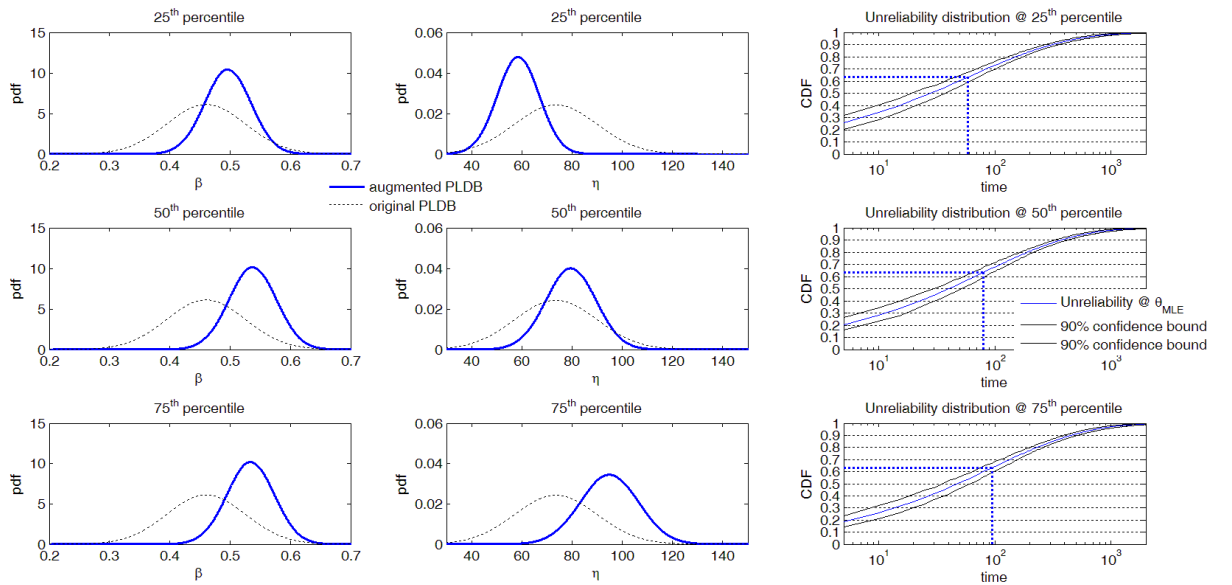

Figure 5: Distributions of the ML estimators in correspondence to different quantiles of the missing data distributions

6 SENSITIVITY ANALYSIS

In Section 5, it has been shown that the methodology proposed in Section 4 to 'reconstruct' the missing inspection times t allows reducing the uncertainty in the component failure time estimates, thanks to the use of the additional 'reconstructed' information.

In this Section, the robustness of the proposed methodology is checked through an analysis of the sensitivity of the uncertainty reduction to the composition of the dataset (e.g., ratio between the cardinality of the merged dataset and that of the original dataset).

6.1 *Proposed Approach*

The approach proposed in this Section is based on a cross-validation technique: the sensitivity of the parameters of the Weibull distribution describing the degradation mechanism of the turbine nozzle is computed with respect to the variability of the number of available inspection outcomes.

In details, we first recall that from the original dataset containing the outcomes of the N=30 inspections performed on machines we have computed:

    a. The width W_η^N of the double-sided 90% confidence intervals for η (W_η^N =55).

    b. The width W_U^N of the double-sided 90% confidence intervals for the unreliability at t= $\hat{\eta}$ (W_U^N =0.13).

These values, W_η^N and W_U^N are used as reference points to evaluate the amount of reduction in the uncertainty. We apply the following procedure:

1. Select randomly from the additional dataset a number M_1 out of M turbine inspection outcomes.
2. For each of the M_1 records, randomly sample the equivalent start from the distribution corresponding to the number of its nozzles found failed at inspection (as explained in Section 4).
3. Re-estimate the Weibull parameters (as in Section 3), on the basis of the data relevant to all the $N+M_1$ inspection times.
4. Estimate the width $W_\eta^{M_1+N}$ and $W_U^{M_1+N}$ of the double-sided 90% confidence intervals for η and for the unreliability at t= $\hat{\eta}$, respectively.
5. Repeat steps 2-6 a large number of times (e.g., 1000).
6. Average the results collected at point 5, and calculate the following Figures of Merit (FOMs):

$$FOM_{W_\eta} = \frac{E\left[W_\eta^{M_1+N}\right]}{W_\eta^N} \tag{1}$$

$$FOM_{W_U} = \frac{E\left[W_U^{M_1+N}\right]}{W_U^{S_1}} \tag{2}$$

    The larger the FOMs are, the smaller is the reduction in the uncertainty on the failure behavior brought by the reconstructed data M_1. Moreover, the standard deviation of the estimator of the FOMs is also calculated. This allows describing the uncertainty in the estimate owing to the limited number of MC simulations, which is summarized by the 68,3% confidence interval (bars in Figures 6 and 7).

7. Procedure 2-6 is repeated for different values of M_1.

Results of the procedure are shown in Figures 6 and 7. In both cases, it clearly emerges that the larger the number M_1, the larger the reduction in the uncertainty of η and of the unreliability. In other words, the added information brought by the M_1 inspection outcomes is more significant the larger M_1 is.
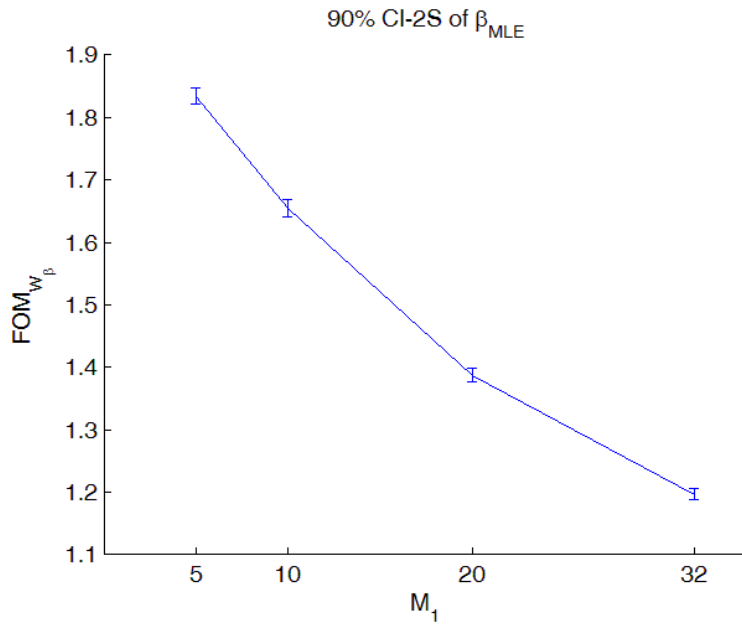
Figure 6 FOM_{W_η} as a function of the number M_1 of available machines
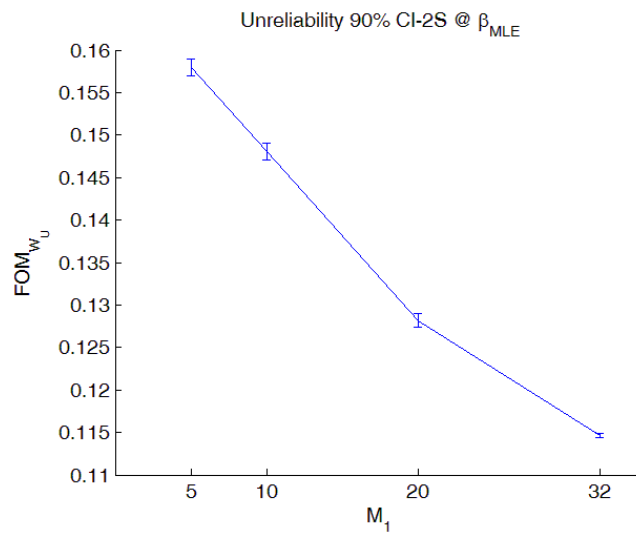

Figure 7 FOM_{W_U} as a function of the number M_1 of available machines

7 CONCLUSIONS

In this work, the procedure presented in (Di Maio et al., 2014) has been discussed and further detailed. This procedure allows to complete the data collected at field inspections of components and, thus, exploit all available data gathered. To illustrate the procedure, the case study of the GE turbine nozzle system has been considered. The procedure is capable of:

– Estimating the parameters of a Weibull distribution as well as the related uncertainties;
– Reconstructing missing equivalent starts to reduce the uncertainty on the estimation, relying on an enlarged database that contains original (complete) data and additional (reconstructed) data.

Finally, it has been shown that the reduction of the uncertainty in the estimates depends on the number of records collected in the field inspections, and the robustness of the proposed methodology has been checked through a sensitivity analysis of the uncertainty reduction to the composition of the dataset.

References

Amato, A., Compare, M., Gallisto, M., Maccari, A., Paganelli, M., Zio, E., 2011, Business interruption and loss of assets risk assessment in support of the design of an innovative concentrating solar power plant. Renewable Energy, 36 (5), pp. 1558-1568.

Baraldi, P., Di Maio, F., Pappaglione, L., Zio, E. and Seraoui R, 2012, Condition Monitoring of Electrical Power Plant Components During Operational Transients, Proceeding of the Institution of Mechanical Engineers, Part 0, Journal of Risk and Reliability, 226 (6) , pp. 568-583.

Baraldi, P., Di Maio, F., Genini, D. and Zio E., 2013, A fuzzy similarity based method for signal reconstruction during plant transients, Chemical Engineering Transactions, 33, pp. 889-894.

Birolini A. 2004. Reliability Engineering, Theory and Practice, 4th Ed., Springer-Verlag.

Brock, G. N., Shaffer, J.R. and Blakesley R.E., 2008, Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. Bioinformatics, pp. 9:12.

Compare, M., Martini, F., Mattafirri, S., Carlevaro, F., Zio E., 2014, Semi-Markov model for the oxidation degradation mechanism in gas turbine nozzles. Submitted for publication to IEEE Reliability Transactions.

Di Maio, F, Compare, M., Mattafirri, S. and Zio E. 2014, A double-loop Monte Carlo approach for Part Life Data Base reconstruction and scheduled maintenance improvement, Safety and Reliability: Methodology and Applications Edited by Tomasz Nowakowski, Marek Młyńczak, Anna Jodejko-Pietruczuk, and Sylwia Werbińska-Wojciechowska CRC Press 2014, pp. 1877–1884.

Hess, A., Calvello, G., Frith, P., 2005, Challenges, Issues, and Lessons Learned Chasing the "Big P": Real Predictive Prognostics Part 1, IEEE Aerospace Conference.

Hines, J.W., Wrest, D.J., and Uhrig R.E., 1996, Plant Wide Sensor Calibration Monitoring, In Proceedings of the 1996 IEEE International Symposium on Intelligent Control, Dearborn, Michigan. IEEE, Piscataway, New Jersey, pp. 378-383. September 15-18.

Kendall, M. and Stuart A., 1979, The Advanced Theory of Statistics, Vol. 2, Charles Griffin and Company limited, London & High Wycombe.

Kim, H., Golub, G.H. and Park H., 2005, Missing Value Estimation for DNA microarray gene expression data: local least squares imputation, Bioinformatics, 21(2):187–98.

Luh, G. C. and Lin C.Y., 2011, PCA based immune networks for human face recognition Applied Soft Computing Journal, 11 (2), pp. 1743-1752.

Meeker, W.Q., Escobar, L.A., 1995, Teaching About Approximate Confidence Regions Based on Maximum Likelihood Estimation. The American Statistician, 49(1), pp. 48-53.

Murthy, D.N.P. and Asgharizadeh E., 1999, Optimal decision making in a maintenance service operation. European Journal of Operational Research, 116, pp. 259-273.

Nelson, W., 1982, Applied Life Data Analysis, John Wiley & Sons, Inc., New York.

Papoulis, A. and Pillai S.U., 2002, Probability, Random Variables and Stochastic Processes, 4th edition, International Edition, Mc-Graw-Hill.

Pipe, K., 2008, 2008, Practical Prognostics for Condition Based Maintenance, proceedings of the Internantional conference on Prognostics and Health Management.

Schafer, J.L. and Graham J.W., 2002, Missing data: our view of the state of the art, Psychological Methods, Vol. 2, pp. 147-177.

Spall, J.C., 2005, Monte Carlo Computation of the Fisher Information Matrix in Nonstandard Settings. Journal of Computational and Graphical Statistics, 14 (4), pp. 889–909.

Zio, E., 2013, The Monte Carlo Simulation Method for System Reliability and Risk Analysis Springer Series in Reliability Engineering.

Zio, E., 2007, An Introduction to the basics of reliability and risk analysis. World Scientific Publishing, Singapore.

Zio, E., Compare, M., 2013, Evaluating maintenance policies by quantitative modeling and analysis, Reliability Engineering and System Safety, 109, pp. 53–65.

Wang, W., 2010, A model for maintenance service contract design, negotiation and optimization. European Journal of Operational Research, 201, pp. 239–246.