Advances in Reliability, Safety and Security, Part 6 – Kołowrocki, Magryta-Mut (eds.)
 ISBN 978-83-68136-18-0 (printed), ISBN 978-83-68136-05-0 (electronic)

Advances in Reliability, Safety and Security

ESREL 2024
Monograph Book Series

Practical Use Of Data Lake To Improve Fatigue Life Estimation

Amaury Chabod

HOTTINGER BRUEL & KJAER France SAS, Sucy-enBrie, France

Abstract

The qualification of components in terms of Durability and Reliability relies on the analysis on a lot of sensors, CAN, IIOT data, which needs a data management infrastructure, to understand the customer usage and variability. Such big data infrastructure is often called « data lake », and may lead to storing huge amount of data. This infrastructure must be generic yet test data-oriented to understand the data structure and its analysis required, and to be optimized for such application. All those data may come from connected equipment, instrumented fleets, test lab or proving ground measurements, digital twins and multi-body dynamics simulations. Further, the data must be managed in terms of quality and traceability. It must be indexed to be able to be retrieved through searches (customer, vehicle, measurement site, engine specification, road condition, usage conditions).
Once this step is achieved, the product development team is able to have a better understanding of customer usage and inputs variabilities in different environments and conditions, which have to be taken into account in the product's mission profile or duty cycle. This ad-hoc mission profile enables the creation of realistic, meaningful design and test specifications.
Understanding the customer usage and input variabilities enables a probabilistic approach to fatigue life prediction. The uncertainties on inputs (geometry, material and loading) may be propagated through the life process, knowing each input's probability distribution function, using a Monte Carlo analysis. The infrastructure enables the life analysis to be done through multiple runs on cloud-oriented server, which enables automation and streamlining the whole process. A use case will be presented to illustrate the approach and its benefits.

Keywords: data ake, uncertainty quantification, Monte Carlo, probabilistic fatigue

1. Introduction

The purpose of this paper is to improve engineers' ability to predict product life by leveraging information in a data lake. Firstly, in view of improving the use of measured loading data, we will describe how to merge data from various sources, clean and organize it for easy access, and build a better understanding of product usage, also known as the mission profile or duty cycle. This data lake infrastructure will be used to quantify the uncertainties inherent in loads, and finally to deploy a probabilistic fatigue process in a comprehensive and automated process.

1.1. Test database

The digital transformation movement is a general trend, affecting many aspects of human activity. Data storage capacities, computational efficiency, and artificial intelligence are making this transformation global. Companies can now store vast amount of data, and build very large databases. The first pillar of big data is Volume, with ever-increasing quantities of data. The use of plural for database illustrates one aspect of this processing: data conditioning is not simple, and shows the second aspect, which is Variety, meaning many types of data. The third aspect, a consequence of the previous two, is the required speed and computer efficiency of the hardware and software needed to assimilate this growing volume. This assertion is summed up under the name of Big Data, with the rules of the 3 Vs: Volume, Variety and Velocity.

With availability and agility of cloud infrastructure, data is not a challenge: the main issue consists to extract value from the data.

The Variety of data requires a pre-conditioning phase, to normalize the form of the data, so that it can be processed on a massive scale. In first place, the data may lie in several areas, network drives, cloud-based proprietary storage folders, and internal databases. Accessing this Variety of data and sharing to different departments (design, CAE, test) traditionally require a deep knowledge and programming skills. A task as fundamental as standardizing channel names after years of acquisition with different sources can be an enormous process. This issue means it is essential to enable a standardization process, or to have the same traceability on channels regardless of their source. Variety also results from the mix of time series measurements from test data acquisition systems and, increasingly, from the CAN bus, where data has a large number of channels and is unevenly time-stamped. To give value to the data, outliers must also be removed, to avoid using unrealistic, values as inputs into engineering designs, simulations and data science.

This phase requires a modern engineering tool optimized to handle signal test data, whereas standard big data tools may not be fully optimized in this area. To achieve this goal, nCode GlyphWorks and nCodeDS signal processing software play this role, as described in the nCodeDS white papers (2019).

The order of magnitude of the data collected is described below as an example of volume in various industries:

- In the wind energy sector, 8,000 files can be collected per year.
- In the automotive industry, connected vehicles (customers and test cars) generate 6 TB per year.
- In the aeronautics industry, 3 GB of data, with 10,000 channels, are collected from avionics per flight.

1.2. Indexing data and making requests

Storing data over years, across multiple projects, design iterations and conditions, requires precise characterization of the context associated with the data. Without context, the value of the data is close to zero, as any comparison, synthesis and understanding is impossible. However, a discussion must take place in order to avoid adding too much contextual information and finding the right balance, and focusing on relevant information. What's more, even if data lies on intranet infrastructure, adding product-specific context may raise security issues, which must be addressed by state-of-the-art security policies and technologies (SSL encryption, SSO). The big data infrastructure requires a tag indexing phase, with :

- Project data: project ID, system ID, part ID, iteration.
- Technical data: engine type, engine power, torque, tires, gearbox type.
- Conditions: test bench, test field, customer use.
- Environment: temperature, weather conditions, type of road.
- Test context: test identification, test reports, documents, videos
- Calculated tags: statistics, metrics

Nowadays, context can be added for ground vehicles with CAN data, which describes all vehicle activity.

There are various databases, with a strong hierarchical structure, such as SQL-type databases. There are also unstructured databases, described as no-SQL databases, such as MongoDB used in nCode Aqira. There's nothing to prevent the use of hierarchical tags (such as a folder tree) and unstructured tags. The advantage of no-SQL databases is greater flexibility, while retaining the ability to maintain a data structure.

Once indexed, raw data is processed and normalized, cleaned and processed signals are indexed as well.

This enables the user to make queries, choosing his or her data basket as in a standard consumer website, and to query on simulation scenarios, including or excluding certain conditions defined by tags (temperature, references, test conditions, etc.). The nCode Aqira infrastructure for big data testing features a web server, installed on the user's intranet, which links all the different aspects of processing, using engineering applications to automate and streamline the process:

- Data cleaning and merging, signal processing for useful metrics and statistics (nCode GlyphWorks)
- Index data on MongoDB database (Aqira)
- Request data (Aqira)
- Post-process and dashboard creation

An example

- The data:

An application was built with 500 files, containing 6 channels, from 12 countries, 4 car models (Berline/SUV/Break

CityDweller), 5 driving conditions (City/Highway/Road/Belgian Block/Offroad) and 4 engine types (Electric/Hybrid/Diesel/Gasoline). For example, the process of requesting a set of 34 files, according to request labels (Model=Berline and Engine=Electric), makes it practical to build a duty cycle, a list of files describing

several events, made of to two time series channels Fx(t) and Fy(t), weighted with a number of repetitions, also stored as a label.

- Searching:

On a web page, the query is easily performed, without programming, as is common on consumer-oriented websites, by ticking tags values, or choosing tag value directly or between boundaries. The resulting file list is reviewed by the requestor, then this list of time-series events according to the number of repetitions is directly pushed as input to the fatigue analysis (Figure 1).

- Fatigue analysis:

For each event, stress fatigue cycle is created by linearly superimposing the stress responses σ(Fi) from finite element analysis to measured load histories Fi: $\sigma(t) = Fx(t)* \sigma(Fx)+Fy(t)* \sigma(Fy)$. Then, an stress-life fatigue analysis is performed, using linear accumulation of damage (Miner's rule), to compute damage for each event and total fatigue damage (HBK (2024)).

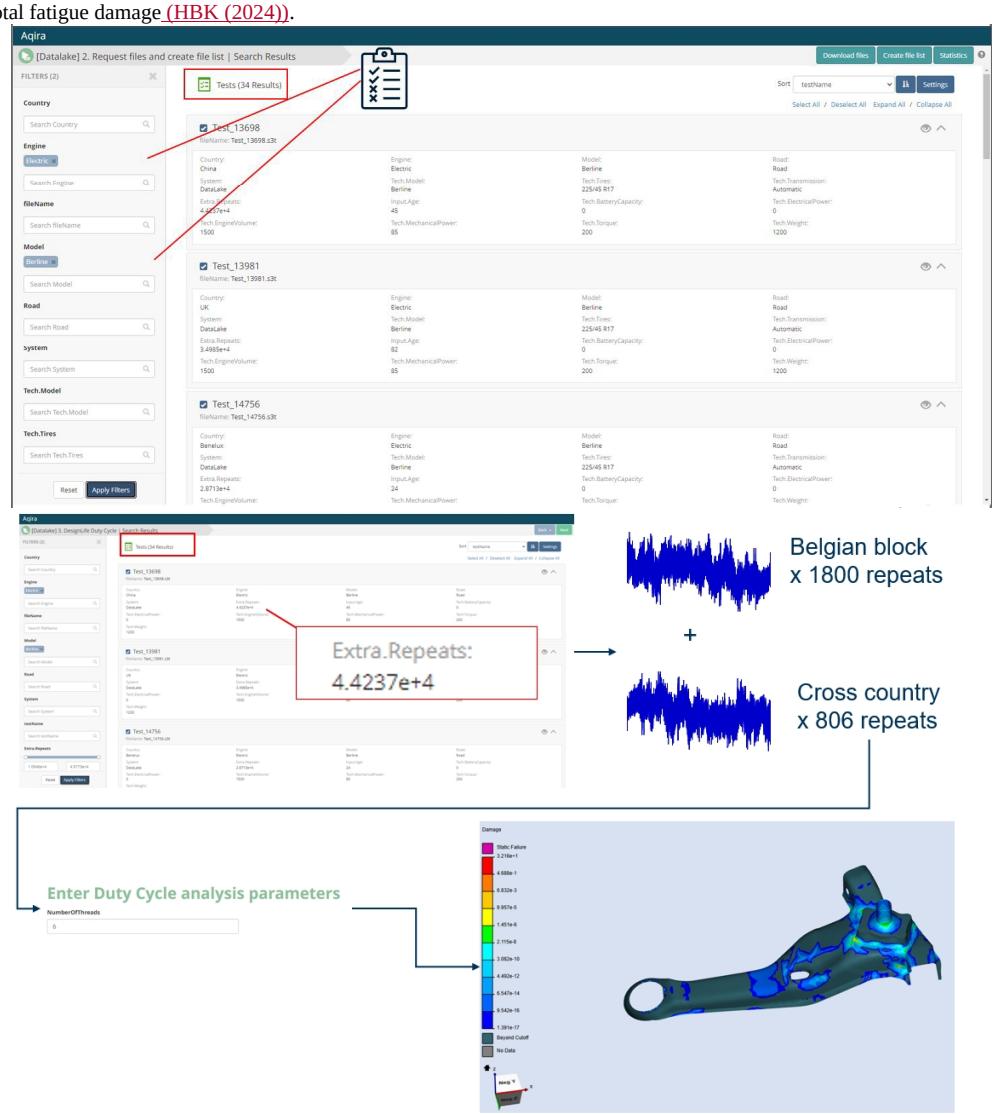

Fig. 1. App that automatically builds the duty cycle from a time series request based on the number of repetitions, and launches a fatigue analysis.

1.3. Probabilistic Fatigue

- Understand uncertainties on loads

The general fatigue route involves a fine-tuned description of the fatigue properties of materials, geometry and loads applied to the component in service. A detailed description of probabilistic fatigue methodology has been elaborated in a previous paper (Halfpenny, Bonato et al. (2019)). The current study is aimed at establishing and democratizing such a process in a datalake environment. There are two natures of uncertainty:

- Reducible uncertainties (or epistemic uncertainties)
- Irreducible uncertainties (or aleatoric uncertainties)

Material uncertainties are inherent to the material's microstructure and therefore irreducible, even with a large number of samples used to characterize fatigue properties. Load uncertainties can be reduced by improving knowledge of customer usage and environmental conditions. This article focuses on load uncertainties derived from customer usage. The massive amount of loading data in the data lake will be used to understand this uncertainty.

One way of characterizing any random variable is the probability density function or PDF, using a distribution such as normal, log-normal, discrete, Weibull, or Rayleigh. The PDF can be applied to characterize the time series loading signal, with the mean value and standard deviation in the case of the normal distribution. Distribution fitting methods may include rank regression or maximum likelihood estimation (MLE), particularly suited to large samples and uncensored data.

The PDF can also be applied to loading stored in a histogram format, instead of considering all the data points in the time signals. Indeed, a clever way of extracting relevant information for fatigue analysis is to construct histograms to build the mission profile, such as rainflow cycle count, FDS, RDS, or time at level, used by Chojnacki et al. (2019).

The usage of a data lake provides a convenient way to search the database, retrieve events and such calculated histograms associated with parent events, and at the end to get mission profile or histogram PDF. As described in NF X 50-144-3 (2021), for each value of the histogram, the damage at each frequency for example, the PDF parameters can be adjusted, and from a percentile value p=1-α, we can derive a design-focused value of FDSα. The way to calculate these FDSα is described in Table 1.

Table 1. Computation of severized histograms.

		Weibull	Normal	Log-Normal	Rayleigh
Probability Density Function	f(t)	$\frac{\beta}{\eta}(\frac{t-\gamma}{\eta})^{\beta-1}e^{-(\frac{t-\gamma}{\eta})\beta}$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}$	$t'=\ln(t)\frac{1}{\sigma'\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{t'-\mu'}{\sigma'})^2}$	$(\frac{t-\gamma}{\sigma^2})e^{-\frac{1}{2}(\frac{t-\gamma}{\sigma})^2}$
Cumulative Density Function	F(t)	$1-e^{-(\frac{t-\gamma}{\eta})\beta}$	$\int_{-\infty}^{t}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}dx$	$\int_{-\infty}^{t'}\frac{1}{\sigma'\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu'}{\sigma'})^2}dx$	$1-e^{-\frac{1}{2}(\frac{t-\gamma}{\sigma})^2}$
Reliability	R(t)	$e^{-(\frac{t-\gamma}{\eta})\beta}$	$\int_{t}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}dx$	$\int_{t'}^{\infty}\frac{1}{\sigma'\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu'}{\sigma'})^2}dx$	$e^{-\frac{1}{2}(\frac{t-\gamma}{\sigma})^2}$
FDSa	t(R=α)	$\gamma+\eta.(-\ln(1-R))^{1/\beta}$	$\mu+\sigma.F^{-1}(R)$	$e^{(\mu'+\sigma'.F^{-1}(R))}$	$\gamma+\sigma\sqrt{-2\ln(1-R)}$

The FDSα is designed to characterize the non-stationarity nature of a signal, using the MBD method, and can also be applied to several events to quantify load uncertainties. If several tests are merged as part of the duty cycle, a time/distance weighted average operation must be performed, before adjusting the PDF parameters for each histogram class.

From this fatigue damage spectrum, design-specific to the target percentile value, a synthetic random PSD can be built. This vibration profile can then be used for analytical life prediction and physical durability validation testing.

The example below shows the FDSα based on several events in Figure 2.

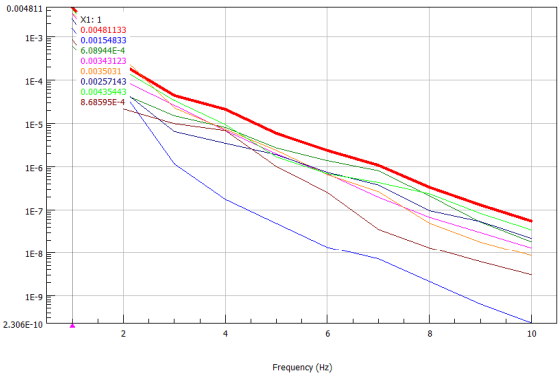

Fig. 2. FDSα (in bold red) obtained from the FDS of 7 events, at 90% percentile value.

In our practical application of probabilistic fatigue, we will focus on characterizing the uncertainty of loading defined by a Gaussian distribution, with a load's known mean and standard deviation, as opposed to the deterministic approach where conservative values are taken according to a fixed percentile value, or an assumed safety factor must be used. The aim is to subsequently generate a Monte Carlo process. The data lake environment provides an easy-to-use engineering app that encapsulates the analysis flows in a web page environment. To characterize the uncertainty of loading, the user selects a basket of files using search criteria like project, technical or calculated labels. The data files meeting these search criteria are sent to an nCode GlyphWorks signal processing flow, which can calculate mean and standard deviation statistics in the background on the server. At the end, the user receives a report in which the most important information is summarized, with the total duration of all files processed (34 files):

Table 2. Summary report of 34 analyzed time series tests (from user request Model=Berline and Engine=Electric).

Title	Value	Units
Duration	7.66E+04	s
Test	34	-

Table 3. Statistical parameters of standard distribution (from user request Model=Berline and Engine=Electric).

ChanNumber	ChanTitle	Mean	SDev
1	X-Force	-9.57	469
2	Y-Force	142.4	469.4

Based on this information, uncertainties are characterized by a normal distribution over these two channels (Fx and Fy), with the parameters defined in Table 3. These statistics describe the variability of two load directions from all events processed, and can now be incorporated as variable inputs in the probabilistic fatigue analysis.

- Monte Carlo process

After quantifying the uncertainties on the inputs (load, geometry or material) with their distributions, a Monte Carlo process is set up, in which a large number of random simulations are run, in order to explore the variabilities of the inputs (Iman et al. (1984), Taguchi et al. (2005), Rubinstein et al. (2016)). Random input values are generated from random numbers between zero and one, using the inverse cumulative density function (CDF).

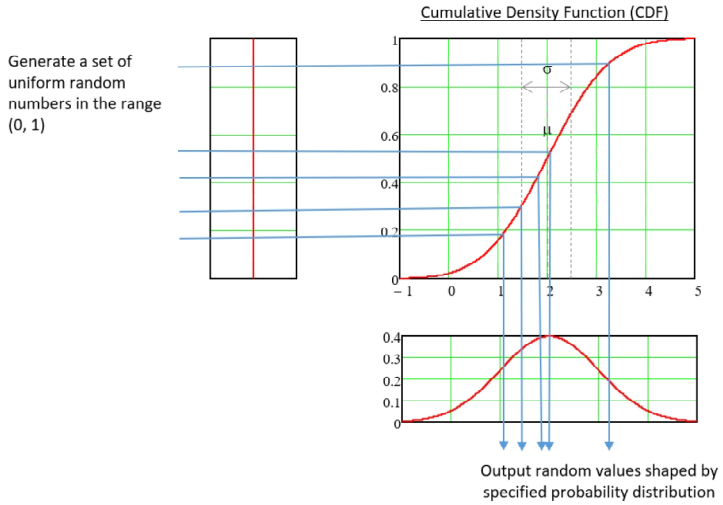

Fig. 3. Overview of the Monte Carlo random number calculation method.

The unitary process consists of repeating the simulation for any set of random input values. The main process is a fatigue analysis, as described below in Figure 4, where the amplitude scaleFx and scaleFy will be defined later as two independent random input parameters. The FE model has two unit load cases Fx=1 kN and Fy=1 kN, and a constant amplitude stress fatigue cycle is created by linearly superimposing the stress responses σ(Fi) to unit loadcases Fi: σ(t) = scaleFx* σ(Fx) on first time step and σ(t) = scaleFy* σ(Fy) on second time step of cycle. In this use case, damage analysis is performed using nCode DesignLife 2022.0. The assumption of linear behavior and high-cycle fatigue leads to the use of the Basquin SN curve and the linear accumulation of damage, in accordance with Miner's rule.

By repeating the simulation for any set of random scaleFx and scaleFy values, according to their PDF form with the corresponding mean and standard deviation previously calculated (see Table 3), uncertainty is propagated into the simulation result, in our case the predicted fatigue life. The number of simulations is 30 for the current application. The advantage of a data lake environment is that the PDF parameter results are adjusted according to the user's request (here: model=Berline and motor=Electric). The single deterministic simulation using the mean values of scaleFx and scaleFy from Table 3 gives a lifetime of B50=5.17e8 cycles.

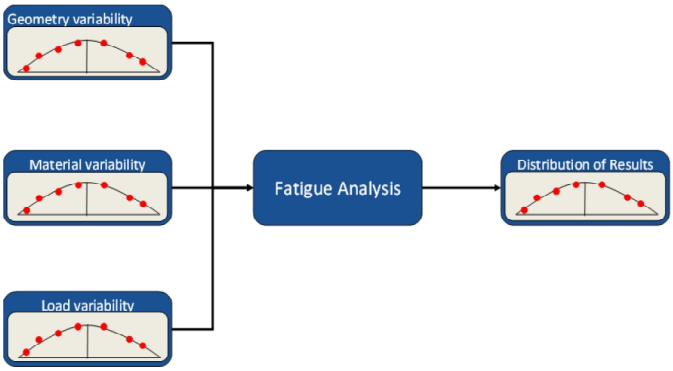

Fig. 4. Overview of the Monte Carlo simulation process.

The advantage of using a data lake is that random values can be easily generated and all batch simulations run efficiently on the server. Any parameter in the analysis stream can be considered a random variable. This doesn't avoid any scripting operations, but the Python routines are integrated into the application process and hidden from

the user, making deployment easier compared to the complexity of a Monte Carlo process. Indeed, an app is easily automated using an external call with a web URL, in which any variable and corresponding value are included. Next, we obtain the values of the damage results for any pair of input values (here Fx, Fy), and for the requested number of simulations, as described in Figure 5:

https://aqiraaedemo.corp.prenscia.com/aqira/api/apprun/93/run?Fx={text}&Fy={text}

Fig. 5. URL address for calling the unit simulation in the Monte Carlo process, with variables Fx and Fy.

The analysis flow is used to select the most damaged node in the fatigue simulation. This information can be extended by failure mode, i.e. by component, to track the results of separate hotspots. The DOE results shown in Figure 6 enable the design space to be reduced by retaining only those random variables that are statistically relevant to the results.

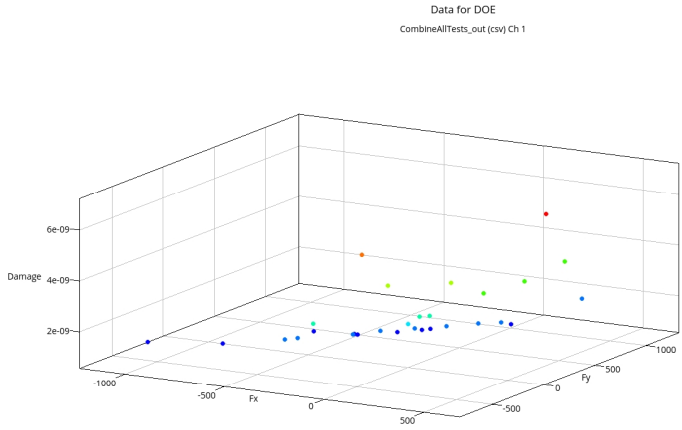

Fig. 6. Simulation DOE showing damage results [-] for input values (Fx [N], Fy [N]).

- Reliability estimation

The final step is to analyze the lifetime values obtained at the most damaged node, which constitute a virtual reliability simulation (ReliaSoft 2015). The PDF and reliability metrics can be analyzed in the same way as in the experimental part, as described in Figure 7.

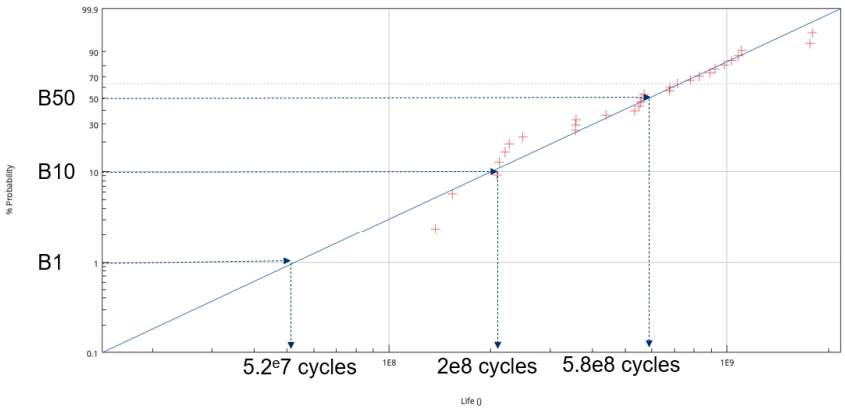

Fig. 7. PDF fitting for Weibull distribution, and median value calculated (B50).

In Figure 7 above, we can plot the median life of 5.8e8 cycles, obtained at 50% probability of failure (B50), for comparison with the previous deterministic value using the parameters of the mean value B50=5.17e8 cycles in paragraph 4.2. This B50 value is mainly used to correlate fatigue failure with the median value obtained from several test prototypes.

The added value of probabilistic fatigue is to obtain a complete description of the variability of service life results with different BX% service life values calculated in Table 4, and Beta and Eta represent Weibull PDF parameters, as defined in Table 1.

Table 4. PDF parameters of Weibull PDF (Beta, Eta), PDF fit Correlation and Reliability Metrics (BX%).

B1%	B10	B50	Beta	Correlation	Eta
5.28E+07	2.01E+08	5.86E+08	1.757943629	95.37453251	7.22E+08

The advantage of running these simulations is to virtually assess the sensitivity of life results on input variability. Unlike the deterministic process, where we check a pass/fail response for a target percentile of users, we can now explore the design space and understand how input data uncertainties affect the dispersion of lifetime results. It produces a virtual reliability test, defined with random samples, and gives the same reliability measures on variability at any failure percentage (B1%, B10%, B50%).

4. Conclusion

A Big Data management and analysis framework is effective and necessary to centralize, standardize, merge, secure access and traceability of test data. It is well suited to understanding load uncertainties and creating a mission profile or duty cycle as a loading specification for a component. For example, the process of requesting, extracting and building a duty cycle from specified conditions is easy to perform, and fatigue analysis can be chained together in the same operation. Finally, the understanding, linking and propagation of uncertainties in numerical simulation can be automated as part of a streamlined process. The Big Data environment makes it possible to link all the steps described, and as a result, the design process is coupled with a data lake to explore numerous scenarios based on user requests, and be able to answer complex product development questions: "Do we need product variants to take account of drastic differences in usage in various parts of the world?".

References

Chojnacki, D., Delattre, B., 2021. Towards A Better Understanding Of Mechanical Stress Applied By Passenger Vehicle Customers With Optimized Instrumentation And Relevant Data Post-Processing Methodologies. Fatigue Design.

Halfpenny, Dr A., Chabod, A., Czapski, P., Aldred, J., Munson, K., Bonato, Dr M., 2019. Probabilistic Fatigue and Reliability Simulation. Fatigue Design.

HBK, 2024. nCode DesignLife theory guide. HBK.

Iman, R., Shortencarier, M., 1984. A fortran 77 program and user's guide for the generation of latin hypercube and random samples for use with computer models. NUREG/cr-3624 sand83-2365. Sandia National Laboratory.

nCodeDS software white paper, 2019. Big Data Analytics. HBK.

nCodeDS software white paper, 2019. Transformational insights from digital bus data. HBK.

Norme NF X 50-144-3, 2021. Démonstration de la tenue aux environnements, Conception et réalisation des essais en environnement, Partie 3 : Application de la démarche de personnalisation en environnement mécanique. Afnor.

Rubinstein, R., Kroese, D., 2016. Simulation and the monte carlo method. Wiley series in probability and statistics. 3rd edition. Wiley.

Reliasoft, 2015a. Experiment design and analysis reference. [Online]. Reliasoft Corporation. Available from: http://www.synthesisplatform.net/references/Experiment_Design_and_Analysis_Reference.pdf.

Reliasoft, 2015b. Life data analysis reference. [Online]. ReliaSoft Corporation. Available from: http://www.synthesisplatform.net/references/Life_Data_Analysis_Reference.pdf.

Reliasoft, 2015c. System analysis reference: Reliability, availability and operation. [Online]. Reliasoft Corporation. Available from: http://www.synthesisplatform.net/references/System_Analysis_Reference.pdf.

Taguchi, G., Chowdhury, S., Wu, Y., 2005. Taguchi's quality engineering handbook. John WILEY & Sons, Inc.