Adaptive response surface method supporting finite element calculations: an application to power electronic module reliability assessment

Carmen Martin, Alexandre Micol, François Pérès

Adaptive response surface method supporting finite element calculations: an application to power electronic module reliability assessment

Carmen Martin [1] · **Alexandre Micol** [2] · **François Pérès** [1]

Abstract In this paper a method is proposed, introducing an adaptive response surface allowing, on the one hand, to minimize the number of calls to finite element codes for the assessment of the parameters of the response surface and, on the other hand, to refine the solution around the design point by iterating through the procedure. This method is implemented in a parallel environment to optimize the calculation time with respect of the architecture of a computational cluster.

Keywords Reliability analysis · Finite element modelling · Response surfaces · Design of experiments · Power electronics

1 Introduction

Finite element modelling is commonly used for the assessment of the reliability of power electronic modules. In recent years, the power of computers has allowed simulating the whole module, taking into account nonlinear aspects of the materials of which it is made. Analytical solutions exist for each individual physical process involved in the fatigue of a module, requiring independent calculations and data exchange between models. The finite element method can be a solution to these issues of process coupling through spatial discretization and resolution of the variational formulation. It is thus possible to troubleshoot electrothermomechanical problems of electronic (Akay et al. 2003; Micol 2007) power modules by a single call to a finite element code. A judicious choice of algorithms, of spatial and temporal discretization accuracy, and of the quality of data must be made to find a sufficiently representative (Suhir 2013; Pérès et al. 2002; Pérés et al. 1999) solution to the problem within an acceptable computing time. This time for calculation is of particular importance when considering the coupling with reliability methods that require a significant number of FE code calls to solve the problem. The use of regression functions to limit the number of requests to the finite element code allows reducing the calculation time. After having briefly recalled in the first part the cornerstone of reliability analysis in the field of mechanics, we introduce here the coupling between reliability and the finite element method based on the use of simple or adaptive response surfaces (Haukaas 2003a; Vallon 2003) and the design of experiments used in their generation . A case study related to power electronic modules, more particularly, IGBT transistors (Insulated Gate Bipolar Transistor), is presented to illustrate these developments.

✉ Carmen Martin
carmen.martin@enit.fr

Alexandre Micol
micol@phimeca.com

François Pérès
francois.peres@enit.fr

[1] ENIT - LGP - INPT- Université de Toulouse, 47, avenue d'Azereix, 65016 Tarbes Cedex, France

[2] PHIMECA Engineering, Centre d'Affaires du Zénith, 34 rue de Sarlième, 63800 Cournon d' Auvergne, France

2 Structural reliability

Predictive structural reliability aims at calculating the probability of failure of a component. A representation or model is used to assess this failure for a given failure mode. The model represents the real mechanical structure as a system with an input, a state and an output (Fig. 1).

For the analysis to be performed, the following parameters have to be defined (Lemaire 2001):

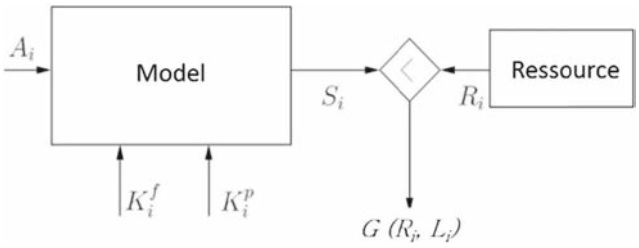


Fig. 1 Resistance load model (Lemaire 2001)

- A_i: mechanical system input, a function of t, generally referring to the load and actions on the system,
- $K_i\ (t)$: mechanical system data split into two categories: K_i^f (data imposed by the specification) and K_i^p (data at the disposal of the designer). They correspond to the geometric features, materials and boundary conditions.
- $R_i\ (t)$: variables related to the capacity of the system to support the load, including the resistance, permissible displacement, and number of cycles before failure.

The structural model allows the simulation of the requirements (load) corresponding to the output of the model, noted L_i If $F(\ldots)$ is the operator of the mechanical model, one can write

$$\mathrm{F}\left(\mathrm{t}, \mathrm{A_i}, \mathrm{K_i^f}, \mathrm{K_i^p}, \mathrm{L_i}\right) = 0 \tag{1}$$

The limit state function allows the assessment of the system robustness by the verification of the inequality

$$G(S_i(t), R_i(t)) \geq 0,\ \forall t \in [0, T] \tag{2}$$

where $[0, T]$ is the required life expectancy or reference period for which the system is studied. In general, the Load Strength Interference model is selected for its simplicity, where function $g(\ldots)$, representing the scenario of failure, is given by the inequality

$$S_i(t) \leq R_i(t),\ \forall t \in [0, T] \tag{3}$$

The system is in the safety region when the calculated load (mechanical model output) remains below its resistance capacity.

Uncertainties need to be taken into account for credible predictions of response of complex systems. Such uncertainties should include uncertainties in the system parameters and those arising due to the modelling of a complex system. The sources of the uncertainties considered in this work are parametric uncertainty - e.g., uncertainty in geometric parameters (friction coefficient, strength of the materials involved, …).

Given $\{X\}$, a random vector made of random variables x_i introducing parametric uncertainty into the mechanical model, function $G(x_i)$, defined by Eqs. (2) and (3), represents the state limit surface when G, vanishes as x goes to 0. $G(x_i) < 0$, defines the domain of failure, and $G(x_i) > 0$ characterizes the safety region. The probability of system failure is thus defined by

$$P_f = \int_{g(\{X\}) \leq 0} f_{\{X\}}(\{x\}) dx_1 \ldots dx_n \tag{4}$$

where $f_{\{X\}}(\{x\})$ is the probability density of vector $\{X\}$.

3 Reliability methods and finite element analysis

It is relatively easy to implement methods of reliability when the failure function G is formally set or known. In many cases, however, the function of failure is not explicit. For realistic structures, the response is calculated by using a numerical procedure such as finite element analysis (FEA) (Jensen et al. 2015). In this case, the derivatives are not readily available, and each evaluation of the function of implicit failure requires a significant computation time. The FORM/SORM (First Order Reliability Method / Second Order Reliability Method) or Monte Carlo methods turn out to be inefficient because of the calculation time required to estimate the function of failure for a relatively large number of points. Within this context, response surface methods can be used to carry out reliability analysis (Micol et al. 2011; Chang et al. 2015; Ben Hassen et al. 2013a; Ben Hassen et al. 2013b) with implicit failure functions.

3.1 Limit surface direct assessment method

3.1.1 Response surface method

The method of the response surface approach leads to the construction of an algebraic expression $\eta: \mathbb{R}^n \rightarrow \mathbb{R}$, most often a polynomial, that is able to approximate the function of failure $H(u)$ or $G(x)$ through its assessment of at different points in the region followed by a regression over these points (Roussouly et al. 2013; Gayton et al. 2003). This method can be inadequate, though, for a completely unknown failure function. The expression of the function of failure obtained from regression analysis is also valid only in the range of the selected random variable values, and extrapolating beyond this interval may be incorrect.

The first step consists, therefore, of selecting a finite number of experimental points from which the value of the function of failure will be evaluated. The most common approach is to take r experimental points by varying the random variable values across several standard deviations $x_i = T_i^{-1}(\pm h)$, where T_i represents the transformation of the x_i variables from the real space to the Gaussian non-correlated space (Alibrandi 2014).

Consider the regression model, linear or not, following observations

$$H_r = H_r(u_k) = \eta(u_k, \theta) + \varepsilon_r, \mathbb{E}_{u_k}(\varepsilon_k) = 0, r = 1, \ldots, N \quad (5)$$

where the vector $U = \langle u_1, \ldots, u_n \rangle$ and η is an unknown function with $\theta = \langle p_0, \ldots, p_m \rangle$ representing the true values of the parameters of the model. The estimation of the parameters $\overline{\theta}$ (characterizing the approximated parameters) requires N observations of $\eta^{(r)} \sim H_r$, $r = 1, \ldots, N$. Measurement errors ε_k are assumed to be independent. Given that the regression model should simulate a mechanical model, $\varepsilon_k(\varepsilon_a, \varepsilon_e)$ includes two types of error. Approximation errors ε_a are errors between the true mechanical model and its approximation by the selected response surface. This error is deterministic and systematic. Experimentation errors ε_e represent the zero mean Gaussian error characterizing the dispersion encountered through the various assessments of H under the same conditions. This error does not appear in the regression of an FE model given that each calculation corresponds to a single result.

However, non-linear FE calculations are flawed by an error of integration, which is not random in time but in space. For the case of drawing observations without replacement for the construction of the numerical design of experiments, it is possible to consider the integration error as an experimental error and assume a zero mean.

For the estimation of parameters $\overline{\theta}$, two methods can be used based on the linearity of function η compared to the parameters to be obtained. η varies linearly with the parameters of the model and can be expressed as

$$\eta(u_k) = p_0 + \sum_{j=1}^{m} p_j \psi_j(u_k) \quad (6)$$

where $\psi_j(u_k)$ represents the regression variables, that is, any linear or non-linear combination of the different components of the input vector U. The various observations can then be expressed as

$$\eta^{(r)} = \eta\left(u_k^{(r)}, \overline{\theta}\right) = p_0 + \sum_{j=1}^{m} p_j \psi_j\left(\left(u_k^{(r)}\right)\right) \quad (7)$$

or also

$$\overrightarrow{\eta} = \Gamma\theta \quad (8)$$

with

- $\overrightarrow{\eta} = \eta^{(1)}, \ldots, \eta^{(N)}$ a vector of N observations of H, that is to say, the exact achievements of the function of failure at the experimentation points calculated in the first step of the method for the input vectors $U^{(r)}, r = 1, \ldots, N$
- Γ the experiment matrix of dimensions $N \times (m+1)$ composed of columns $\Gamma_i = (1, \psi_1(u_k^{(i)}), \ldots, \psi_m(u_k^{(i)}))$

The minimum number of points required for the construction of the response of the approximation surface is then $N_{min} = (m+1)$. In this case, the experiment matrix is square, and obtaining its coefficients can be performed through an inversion of the experiment matrix. The existence of errors in the output of the mechanical model recommends taking a larger number of points and applying a least squares method. The $\overline{\theta}$ estimator is then defined by the minimization of the error function

$$H(\theta) = \|\eta - \Gamma\theta\|^2 \quad (9)$$

and, from the definition of the pseudoinverse,

$$\overline{\theta} = \left(\Gamma^T\Gamma\right)^{-1}\Gamma^T\eta \quad (10)$$

If the function η is not linear with respect to the parameters, an iterative solution of non-linear regression should be considered. Nevertheless, in some cases, the assumption of the parameter non-linearity can be relaxed by not expressing H with respect to $\psi_j(u_k)$ but as a function of $\log \psi_j(u_k)$.

3.1.2 Concept of adaptive response surface

As it has just been shown, the principle of the response surface method is based on the definition of a limit state function to avoid the call to a finite element code. However, several problems may arise in this type of resolution, mainly related to (i) the quality of the response surface and (ii) the effectiveness in terms of computing time. In this work, we will address these shortcomings and will introduce a breakthrough by implementing the concept of the adaptive response surface.

It will be discussed in particular how the different calls to the FE code can be optimized to obtain the results of a reliability calculation. The best strategy for the calculation in the context of non-linear thermo-mechanical fatigue analysis might be to use direct methods, based on the calculation of gradients and Hessians, (Kanchanomai et al. 2002; Van Driel et al. 2003) which seem suitable for performing reliability optimization. These methods appear to be inappropriate, though, when the solution of the FE calculation is vitiated by an error and leads to erroneous gradients. In this respect, the response surface appears to be the best solution to obtain, for all observations, an average behaviour of the FE output and easy calculation of the gradients. The coupling procedure between the mechanical model and the reliability algorithm is shown in Fig. 2.

In the context of parallel computing, the concept of the adaptive response surface that we developed (Micol et al. 2008) enables us to take full advantage of the opportunities

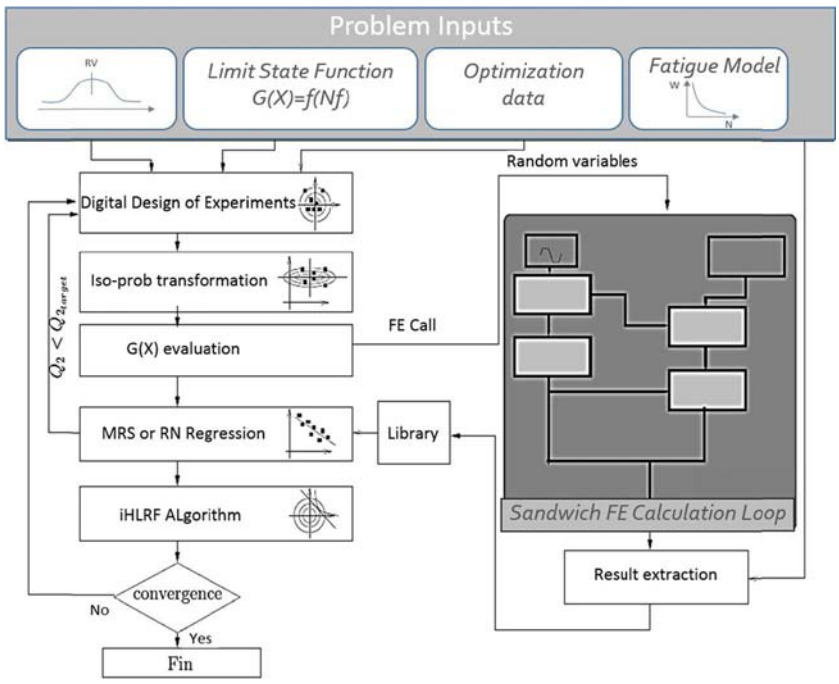


Fig. 2 Coupling between reliability algorithm and finite element calculation

of reliability calculation on a machine cluster. The corresponding algorithm is given in Table 1.

A predetermined type of response surface is first selected as the most efficient one to approximate the output of the calculation. The quality of the response surface requires, then, to minimize the error between the response surface and the real FE evaluation.

The observations to be performed through a design of experiments are planned to maximize the information to be brought. This planning is often based on the minimization of the predictive variance associated with the model for a given experimental variance with the lowest number of experiments. Within the context of numerical simulation, this can be partially omitted because the number of observations influences only the computing time.

Performing the experiments can be achieved in three different ways:

1. Creating a mesh on the experiment region and evaluating the output of the mechanical model for each node. Several types of mesh can be defined according to the type of design of experiments (full factorial, Fisher, Box-Behnken, Taguchi…). These methods have the advantage of scanning the entire domain, but there will be then regular areas not explored by the draw method, and the behaviour of the output will remain unknown.
2. Using the Monte Carlo method to avoid the problem of having completely ignored areas for a sufficient number of draws. In return, some areas may be overestimated by the potential presence of a cluster of relatively close points. Some points then lose their informative quality, while other areas are still underexplored. It is possible to avoid this by analysing the $\boldsymbol{H}$ matrix, which depends only on the planning and may be calculated before the observations. The concept of a leverage point regression then provides information on the observations having an important influence. The values of the trace of $\boldsymbol{H}$ must be as close as possible. If a value $\boldsymbol{H}_{ii}$ does not match the others for the ith observation, its input values may be reconsidered.
3. Perform a Latin hypercube sampling for which a discretization of the domain is carried out and a single observation is made on the hyper-surface containing the observation. Zou (Zou et al. 2002) furthermore, performed a random draw in the possible subdomain of the assessment.

For the non-linear case, a linearization of the experiment matrix is made around the vector of parameters θ:

$$\Gamma = \left(\frac{\partial \eta(u,p)}{\partial p_1}, \ldots, \frac{\partial \eta(u,p)}{\partial p_n} \right)$$

This linearization can be properly achieved only with prior knowledge of the p parameters. Consequently, the different iterations aiming at identifying the design point use the previous response surface parameters to assess the experiment matrix. The first iteration then uses the vector θ to start the optimization process. Special attention must then be given to the starting value identification. Insofar as possible, a first linear evaluation can be

Table 1 Adaptive response surface algorithm

Syntax
← : variable affectation
↵ : variable added into a list
$U^{(i)}$: ith vector of the U vector list
$U_j^{(i)}$: jth component of vector U$^{(i)}$
$\eta^{(r)}$: response surface vector for the rth observation

Functions
plan(X): draw of a uniform random vector with mean X between both values defined in *range*
len(X): length of list X
Gauss(u,s): draw of a random Gaussian variable with mean u and standard deviation s
regression: response surface regression
regression_analyse: variance analysis
Gfun: state limit function evaluation
FORM: FORM Analysis
SORM: SORM Analysis

Data
Q^2_{target}: Q^2 value
range: range of the domain to be scanned
maxiter: maximum number of loops
ε: stop condition
nrv: number of random variables
param: number of parameters to be evaluated

Results
Reliability index β, α vector, sensitivities, elasticities, Q^2

Starting conditions
Q^2 =0
m=iter=loop=1

```
Begin
  While {(β_{k+1}-β_k/β_{k+1}) < ε and m < maxiter}
    β_k ← β_{k+1} ;
    While {Q² < Q²_target}
      If {m=1}
        If {iter=1}
          For {k=1 to cpu ×(param/cpu +1)}
            U^(m) ↵ plan ([0,...,0]^T, range) ;
        Else
          For {k=1 to cpu}
            U^(m) ↵ plan ([0,...,0]^T, range) ;
      Else
        If {iter=1}
          For {k=1 to len (U^(m-1))}
            If {| U_k^(m-1) - u* | <1/m}
              U_k^(m) ← U_k^(m-1) ;
              η_k^(m) ← η_k^(m-1) ;
            Else
              U_k^(m) ← plan (u*, range/m) ;
        Else
          For {k=1 to cpu}
            U^(m) ↵ plan (u*, range/m) ;
      X^(m) ← T^-1 (U^(m)) ;
      η^(m) ↵ Gfun(X^(k)) ;
      θ ← regression (η^(m), X) ;
      R², R²_adjust , Q² ← regression_analyse (θ, η^(m), X) ;
      iter ← iter+1;
    β_{k+1} , u* ← FORM (η, X) ;
    m ← m+1;
  SORM
End
```

performed by setting the function parameters to values that would allow the response surface to be calculated from Eq. (7). The returned values are then used in a first non-linear identification for the calculation of the experiment matrix. A uniform design of experiment (Jin, 2004) is used based on several optimization criteria such as

- maximization of the distance d between the points defined by,

$$d\left(u^{(i)},u^{(j)}\right)=\left[\sum_{k=1}^{n}\left|u_k^{(i)},-u_k^{(j)}\right|^t\right]^{1/t},\quad t=1\ ou\ 2$$

- maximization of the Shannon entropy by maximizing the determinant of the covariance matrix related to the experiment matrix,
- minimization of anomaly $D_a(u)$, defined as the difference between the empirical distribution function of the design of experiment $F_n(t)$ and the distribution function of the uniform law $F(t)$ on the whole domain defined by plane $\mathbb{D}^n$

$$D_p(u)=\left[\int_{\mathbb{D}^n}|F_n(t),-F(t)\ |^p dt\right]^{1/p}$$

Most planning algorithms are based on exchange algorithms because this type of approach is generally

associated with experimental design of experiments, for which the variables are often discrete. Nothing prevents the analyst from restricting this type of variable input into this type of algorithm because the FE code uses continuous values (truncated) as inputs. We can therefore use an optimization algorithm to achieve continuous and bounded input observation planning. The proposed method is then expressed by the optimization problem

$$\begin{aligned}\mathscr{L}(u,\lambda) = &-\min_{u^{(i)},u^{(j)}\forall i,j}\left(d\left(u^{(i)},u^{(j)}\right)\right)\\ &+\lambda_0\left[\boldsymbol{H}_{max}-Tr\left(\boldsymbol{H}-\boldsymbol{I}\frac{N}{m+1}\right)\right]\\ &+\sum\nolimits_i \lambda_i\left[u_{max}-\left\|u^{(i)}\right\|\right]\end{aligned} \tag{11}$$

where

- $\min_{u^{(i)},u^{(j)}\forall i,j}\left(d\left(u^{(i)},u^{(j)}\right)\right)$ is the function maximizing the distance between points,
- $\boldsymbol{H}_{max}-Tr\left(\boldsymbol{H}-\boldsymbol{I}\frac{N}{m+1}\right)$ is the constraint preventing lever points from appearing. The $\boldsymbol{H}_{max}$ value is then to be set as the maximum tolerated with respect to the sum of the differences between the $\boldsymbol{H}_{ii}$ values and the $\frac{N}{m+1}$ mean,
- $u_{max}-\|u^{(i)}\|$ is a constraint related to the norm of each observation to confine them in a circle of radius u_{max}.

The optimization challenge lies in the discontinuity of the performance function. Indeed, the minimum distance between a given point and the others is linear over the displacement of this point in the domain. On the other hand, when maximizing this distance, it has to be noted that a point deviating from its nearest point inevitably becomes closer to another. Therefore, at the location where the minimum distance shifts to another point, a discontinuity in function $\min_{u^{(i)},u^{(j)}\forall i,j}\left(d\left(u^{(i)},u^{(j)}\right)\right)$ appears. This problem is then solved by processing an algorithm, not using the performance function gradients. Optimization is therefore carried out using the algorithm COBYLA (Constrained Optimization BY Linear Approximations (Powell 1994)), which performs a linearization of the gradients in a region of confidence.

Following a uniform drawing of the design of experiment points, the algorithm optimizes the localization of the different points of the plane one after another by assessing, at each step of the optimization, (i) the distance to the nearest other point in the plane, (ii) the experiment matrix Γ to calculate the matrix $\boldsymbol{H}$ and (iii) the distance from the origin. When the first item is optimized, the second is moved to satisfy Eq. (11), and so on:

$$\begin{matrix}\rightarrow\\ x\\ x\\ x\end{matrix}\begin{bmatrix}u_1^{(1)} & \cdots & \cdots & u_k^{(1)}\\ \vdots & \ddots & & \vdots\\ \vdots & & \ddots & \vdots\\ u_1^{(n)} & \cdots & \cdots & u_k^{(n)}\end{bmatrix}\Rightarrow\Gamma=\begin{bmatrix}\psi_1\left(u^{(1)}\right) & \cdots & 1\\ \vdots & \ddots & \vdots\\ \psi_1\left(u^{(N)}\right) & \cdots & 1\end{bmatrix} \tag{12}$$

$$\begin{matrix}x\\ \rightarrow\\ x\\ x\end{matrix}\begin{bmatrix}u_1^{(1)} & \cdots & \cdots & u_k^{(1)}\\ \vdots & \ddots & & \vdots\\ \vdots & & \ddots & \vdots\\ u_1^{(n)} & \cdots & \cdots & u_k^{(n)}\end{bmatrix}\Rightarrow\Gamma=\begin{bmatrix}\psi_1\left(u^{(1)}\right) & \cdots & 1\\ \vdots & \ddots & \vdots\\ \psi_1\left(u^{(N)}\right) & \cdots & 1\end{bmatrix} \tag{13}$$

The algorithm makes several runs on the set of points until the stabilization of their coordinates. To illustrate this method, Fig. 3 shows the case where a full design of experiment would be used for the planning of observations with two basic variables for the regression of a complete quadratic function (including interactions). It appears clear that some points of the design of experiment will then have a great influence on the estimation of the parameters.

After optimization, Fig. 4 gives the location of the different points within, displayed in blue, the perimeter of the considered domain. Two values of $\boldsymbol{H}_{max}$ were used to measure its influence. For a low $\boldsymbol{H}_{max}$, there is no lever point in the identification of the parameters. On the other hand, if the constraint on $\boldsymbol{H}_{max}$ is released, the maximization of the distance between the points is predominant. This allows, then, playing on the constraint of the lever points based on the a priori knowledge of the behaviour of the output to be represented by the response surface. Indeed, if the type of response surface is known, the constraint of $\boldsymbol{H}_{max}$ can be hardened with the risk of having unexplored areas. On the other hand, without any information, it will be better to let the algorithm place points across the domain, even if it means more observations because the Q^2 quality index will inevitably fall.

The planning method of the different basic variables for observations of the design of experiment is first based on a Monte Carlo uniform sampling. The draw will therefore be restricted to a limited domain that the designer will have to set with the aim of including the design point. For a large number of variables, the idea of the location of this point quickly becomes intangible, and the method should be able to move in the domain $\mathbb{D}^n$. After a sufficient number of observations to make null the number of degrees of freedom of the regression model, a first assessment of the predictive Q^2 quality index can be made. As long as this index does not reach the value expected, the realization of new experiments is necessary. In a parallel computing environment, the number of additional observations is equal to the number of available processors. One can thus control the convergence of Q^2 for each calculation time step. A calculation time step corresponds to the time necessary for the FE calculation to end but, in the case of parallel computing, a calculation time step corresponds to as many observations as processors. The number of FE

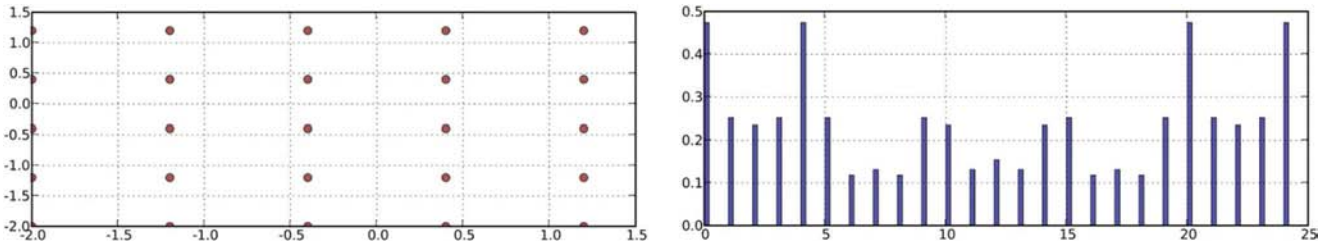


Fig. 3 Full plane pattern and lever point representation

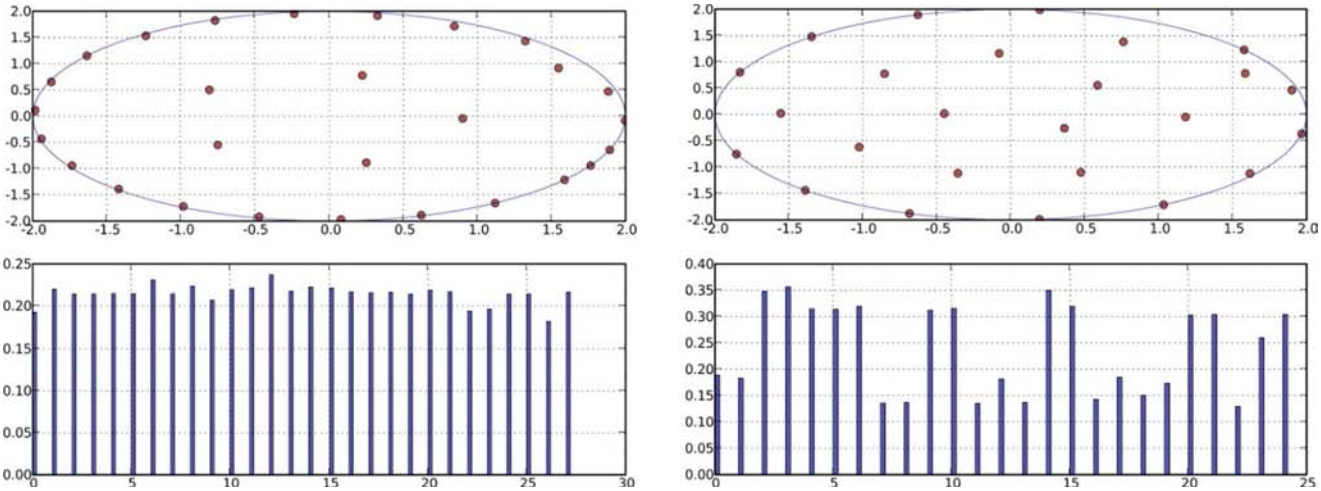
calculations required to achieve a good response surface is then controlled by the available computing power.

Once the index is acceptable (that is Q^2 is close to 1) the evaluation of the parameters of the regression model may be performed together with a first optimization, although imprecise, of the design point. A new regression iteration is then carried out around the design point. The draw is based on the Monte Carlo method with normal distribution whose standard deviation diminishes over the iterations.

The aim is to focus on the assessment of points around the assumed design point. The points of the previous plane within the vicinity of the design point are reused and supplemented by Monte Carlo draws to achieve a sufficient number of points to perform a regression calculation and obtain a correct Q^2. The same concept of adding a number of observations equal to the number of processors is adopted in the subsequent iterations. There are consequently two loops to, on the one hand, reach a good prediction of the regression model and, on the other hand, to move and scale the response surface around the design point.

There are many benefits of using this approach:

– The method is extensible: If a design of experiment does not permit the achievement of a correct Q^2, additional observations are necessary. The algorithm is therefore able to add points in the design of experiment without creating any lever point. The already simulated points are then reused and taken into account in the evaluation of matrix $\boldsymbol{H}$, and only the last points are optimized according to the problem (Eq. (11)). Figure 5a shows the addition of 14 experiments when initially 35 experiments had been carried out and properly optimized. Adding experiments, randomly drawn in the plane, redistributes the different values of the diagonal of $\boldsymbol{H}$. Figure 5b, representing the different values of the diagonal of $\boldsymbol{H}$ after optimization, shows the aptitude of the algorithm to properly incorporate the additional experiments.
– Constraints can be of various types. A constraint in the distance to the origin has been built here to restrict the domain (due to the nature of a failure probability that diminishes when moving away from the centre). However, relational constraints between different variables can also be introduced to take related physics into account. Thus, inequalities modelling mechanical inconsistencies between variables can be considered by transforming them within a normed space and adding them to the optimization problem (Eq. (11)).
– Control of the quality of the model by optimization of the points of the design of experiment (Eq. (11)) is based on the use of a quality index. On the other hand, the iteration

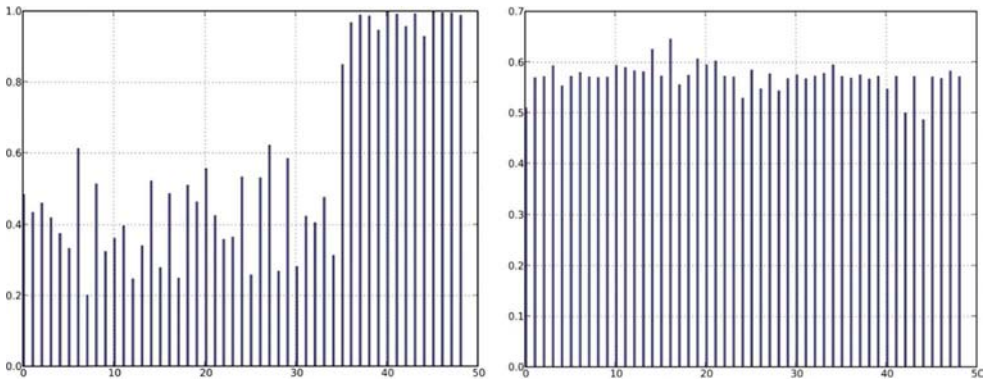


(a) Constraint on lever points high dispersion

(b) Constraint on lever points low dispersion

Fig. 4 Design of experiment optimization

Fig. 5 Addition of points in the design of experiment



(a) Random draw of additional experiments (b) Optimization of additional experiments.

of the construction of a reduced and focused design of experiments allows the representation of the limit state function by a polynomial quadratic function. Limiting the validity of the response surface around the design point reduces the complexity of the real limit state function.

– Parallelization of the optimization code. For the design of experiment planning, the algorithm processes the various points one by one to place them in the right spot within the domain. The code can then be parallelized so that each node of the computational cluster handles a particular point by sending its coordinates to the other nodes at each iteration. The optimization of a point is thus performed through the knowledge of the entire design of experiment in real time.

$$\begin{matrix} \rightarrow \\ \rightarrow \\ \rightarrow \\ x \end{matrix} \begin{bmatrix} u_1^{(1)} & \cdots & \cdots & u_k^{(1)} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ u_1^{(n)} & \cdots & \cdots & u_k^{(n)} \end{bmatrix} \Rightarrow \Gamma = \begin{bmatrix} \psi_1\left(u^{(1)}\right) & \cdots & 1 \\ \vdots & \ddots & \vdots \\ \psi_1\left(u^{(N)}\right) & \cdots & 1 \end{bmatrix} \tag{14}$$

– The method is generic in the sense that a quadratic polynomial function can be sufficient in most cases to approach the limit state function. The iteration of the regression around the design point procedure and the resulting curvature in this reduced domain allows the substitution of a function of a complex non-linear nature with a quadratic function.

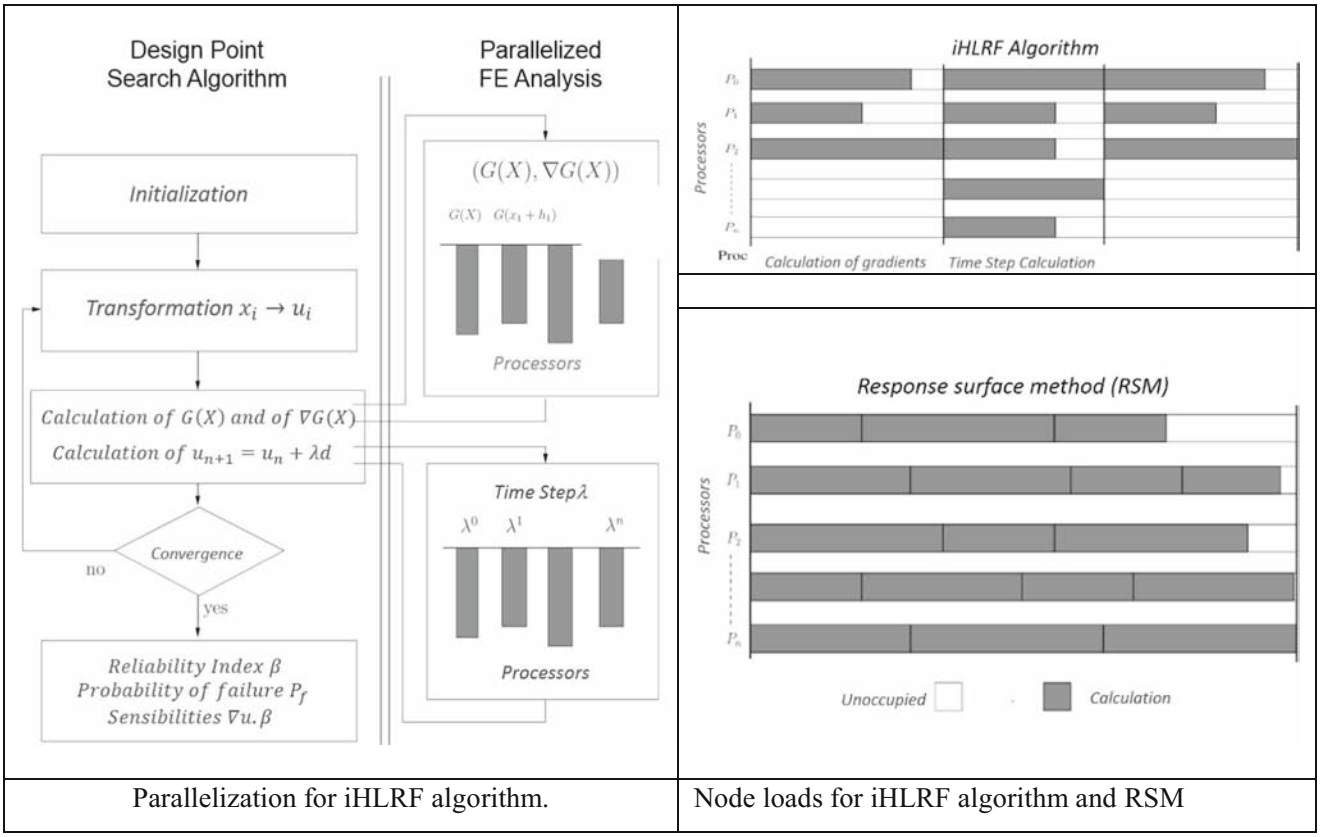
4 Parallelization

As introduced previously, the development of numerical methods is focused on the distribution of tasks to be performed on machine clusters. A few methods stated above can harness the power of parallelization to make it proportional to the number of machines involved. All methods built on regression, neural learning or Monte Carlo simulation require repeating the same loop of FE calculations several times with independent inputs. The computational load can then be stacked and distributed on different processors. Some input data for the FE calculation can be conditioned by previous calls (this would be the case if a regression had to be performed every time a machine changed a value and the next draw was made around the new design point). For the optimization algorithms used for the gradient calculation of each random variable, parallelization allows calculating the finite differences of each variable at the same time. On the other hand, the algorithm is unable to move forward until all gradient calculations are finished.

Figure 6 shows the loads of the various nodes in a cluster in the case of a traditional FORM calculation and those through the use of the response surface method. To obtain the design point with the iHLRF improved algorithm (Lind and Hasofer (1974), Rackwitz and Fiessler (1978)), the gradients of the limit state function are evaluated by finite differences which requires, for n random variables, $(n+1)$ calculations. The time step of the iHLRF algorithm is performed by the Armijo rule, which consists in reducing by half the step if the limit state function is not smaller at point u_{n+1} than at point u. In a parallel environment, the different time steps are evaluated at the same time, and the one that minimizes the merit function is chosen as the new time step.

FORM/SORM approximation methods are then penalized if

– the machine park is not homogeneous (some machines are faster than others) or the FE calculation time of the limit state function differs (a different value for a random variable can lead to a different number of iterations). Some nodes in the cluster then wait until the end of all the calculations to proceed to the finite difference assessment of the gradients.
– the number of CPUs exceeds the number of random variables. This poses no problem in the time step evaluation



Parallelization for iHLRF algorithm. | Node loads for iHLRF algorithm and RSM

Fig. 6 Organization of the calculations for resolution in a parallel environment

because all processors are solicited. For the gradient calculation, however, some nodes will remain unoccupied.

The response surface method does not suffer from these drawbacks because once a calculation is finished, another starts. The load is then 100 % of that possible for the cluster. On the other hand, the response surface method will be less accurate than the direct algorithm due to the approximation made by the response surface. The two methods can be combined in the case of a strongly nonlinear limit state function by first using the response surface to approximately locate the design point and then switching on direct resolution to refine the solution.

5 Case study

To illustrate the proposed response surface method and demonstrate its capabilities, two examples are here presented, using an explicit limit state function in the first one and a call to the FE code in the second one.

5.1 Context

The system studied is based on modules, elements of a railway traction system ensuring a function of power switches. Several technologies based on diodes, transistors and thyristors are available for power modules depending on the power and desired frequency range. The IGBT (insulated gate bipolar transistor) (Ciappa 1997; Jayant Baliga 2015; Liu et al. 2011; Pedersen et al. 2015) studied here is a closing and opening controllable transistor). It combines the advantages of the MOS transistor, being quick and easy to control, and of the bipolar transistor, able to support high voltage levels but also to lower the voltage in the presence of high currents. We focus, in this study, on the design of new power module structures developed in our laboratories and based on flip-chip microelectronic technologies (Fig. 7). We call such devices “NT modules”. The principle is to assemble two substrates to form a ‘switch’ in a sandwich structure (Fig. 8) and replace “wire-bonding” connections with “bump” type connections.

The “bump” connection is made up of three elements: two tin-brazed solders on both sides of a nickel-plated copper insert. This connection ensures three functions (Micol et al. 2009; Darveaux & Banerji 1992; Wang et al. 2001). The first is mechanical because the integrity of the structure is obtained by these connections, both substrates being maintained by the “bumps”. The second is thermal, with better dissipation of heat, which can occur through both sides of the substrate. The third is electrical, as the power is brought to the component through these connections (Lhommeau et al. 2007; Nelhiebel et al. 2013; Castellazzi and Ciappa 2008; Van der Broeck et al. 2015; Deshpande & Subbarayan 2000; Feller et al. 2008).

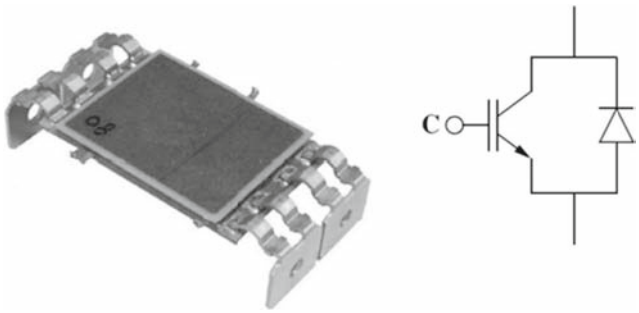


Fig. 7 Prototype of elementary switch and associated circuit diagram

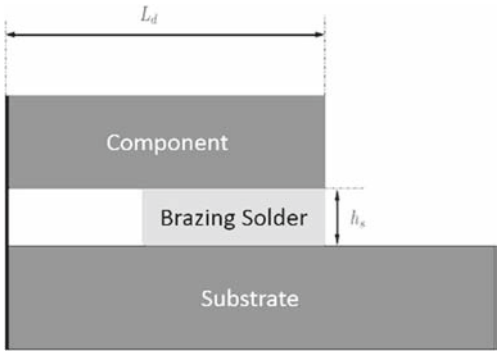


Fig. 9 Simplified structure of component

5.2 Method implementation

Example 1 - The first case addresses the assessment of a brazing solder (Fig. 9) using the fatigue analytical model of Engelmaier. The aim here is to highlight the relevance of the quality indexes $R^2, \overline{R^2}$ and Q^2 and the use of the iteration loop of the response surface procedure.

The strain calculation within the solder can be analytically expressed as (Engelmaier, 1991)

$$\Delta\gamma = C\frac{L_d}{h_s}\Delta\alpha\Delta T$$

where C is an empirical correction factor, L_d the half-component dimension, h_s the thickness of the solder, and Δa and ΔT the Coefficient of Thermal Expansion (CTE) and temperature difference between the component and the substrate, respectively. The energy dissipated within the solder is then calculated by

$$\Delta W = \Delta\gamma\tau$$

with τ the shear constraint within the solder (Lau et al. 1997).

The law of fatigue of Engelmaier assesses the number of cycles before failure through the following equation (Syed, 2004):

$$N_f = (0.0015\Delta W_{acc})^{-1}$$

The variables for this example are presented in (Table 2).

The limit state function characterizes the failure for components not reaching a given number of cycles $F_{f_{target}}$:

$$G(x) = \left(0.0015\left(C\frac{L_d}{h_s}\Delta\alpha\Delta T\right).\tau\right)^{-1} - N_{f_{target}} < 0 \textit{ in the failure domain}$$

The assessment of the number of cycles before failure for the mean values of random variables results in a number of cycles $N_f = 2469$. A resolution of the problem of reliability was made on this analytical model by using the FORM method, leading to a reliability index of $\beta = 3.42 (P_f = 0.00031)$. This calculation was followed by a Monte Carlo resolution method with an important draw around the design point estimated by the FORM method. The result corresponds to a value for the probability of failure of $P_f = 0.00027$, corresponding to a reliability index $\beta = 3.46$.

The real function of failure is here non-linear with respect to the parameters. Suppose this function is unknown and comes from an output of a FE calculation. Without any information on the form of the response, one can decide to approximate it by a quadratic response surface, which represents 15 variables to identify. Three methods for obtaining the design of experiment are used here: (i) the traditional method of establishing a grid on the domain (4 variables × 5 levels/variables = 20 experiments), (ii) for the same number of points, a random draw of observations following a uniform distribution, and (iii) an

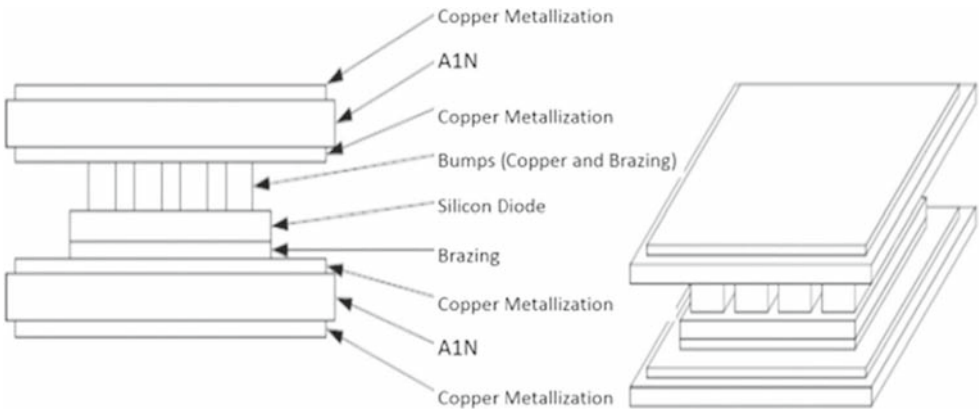


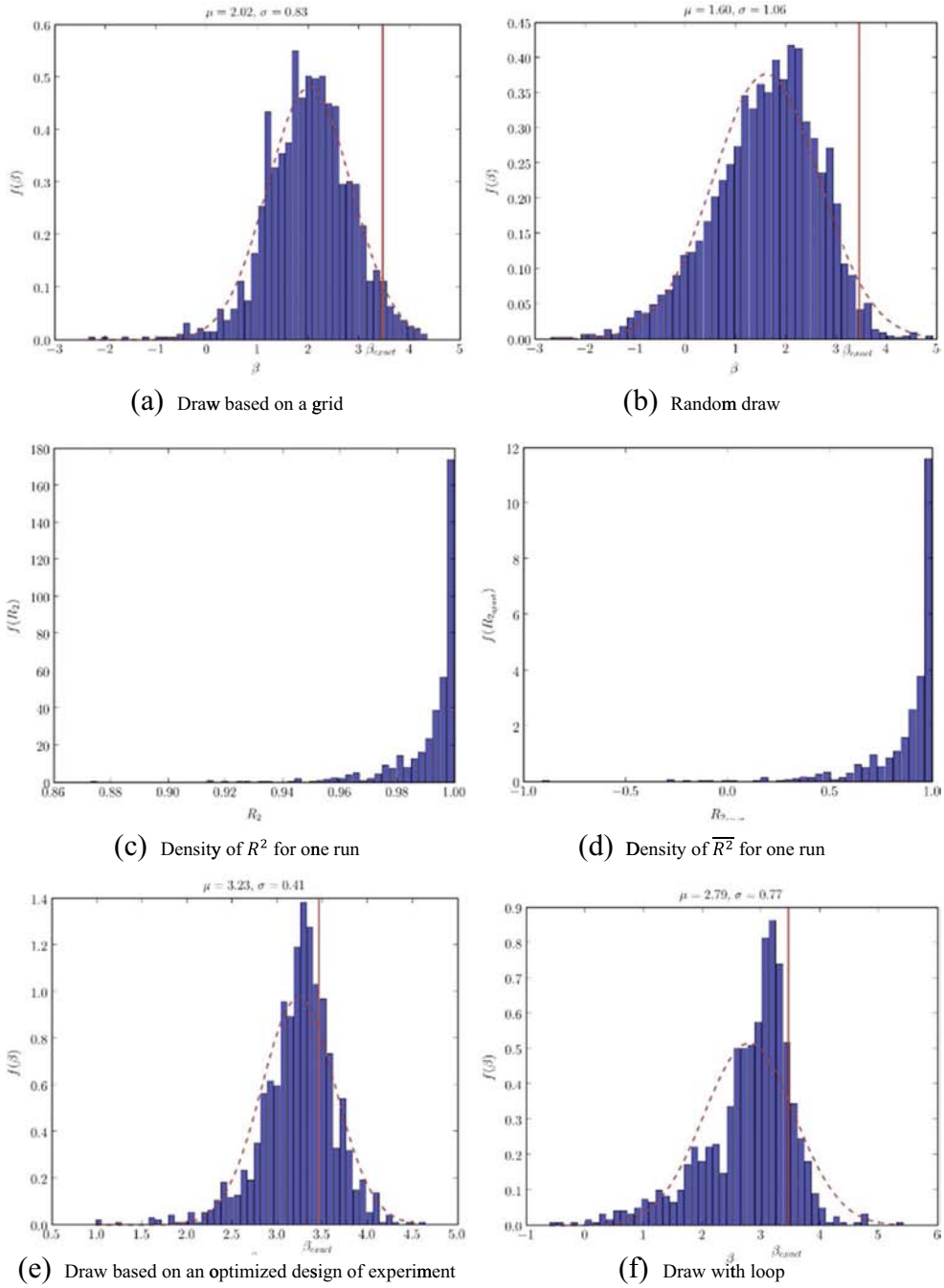
Fig. 8 Basic switch components

Table 2 Variables of the SMC component reliability

Variables	C	L_d	h_s	$\Delta\alpha$	ΔT	τ	$N_{f_{cible}}$
Type	Gaussian	Fixed	Gaussian	Fixed	Gaussian	Fixed	Gaussian
Mean	0.5	3	0.1	5.10^{-6}	180	20	1000
Standard deviation	0.5	–	0.01	–	30	–	200

optimized design of experiment. The evaluation of the limit state function is obtained by adding a Gaussian random variable $N(0,200)$.

Figure 10a and b show the reliability index distribution for the cases where the points of the design of experiment form a grid on ±3 standard deviations and where the



(a) Draw based on a grid

(b) Random draw

(c) Density of R^2 for one run

(d) Density of $\overline{R^2}$ for one run

(e) Draw based on an optimized design of experiment

(f) Draw with loop

Fig. 10 -β and R^2 density values with respect to the method of production of the design of experiments

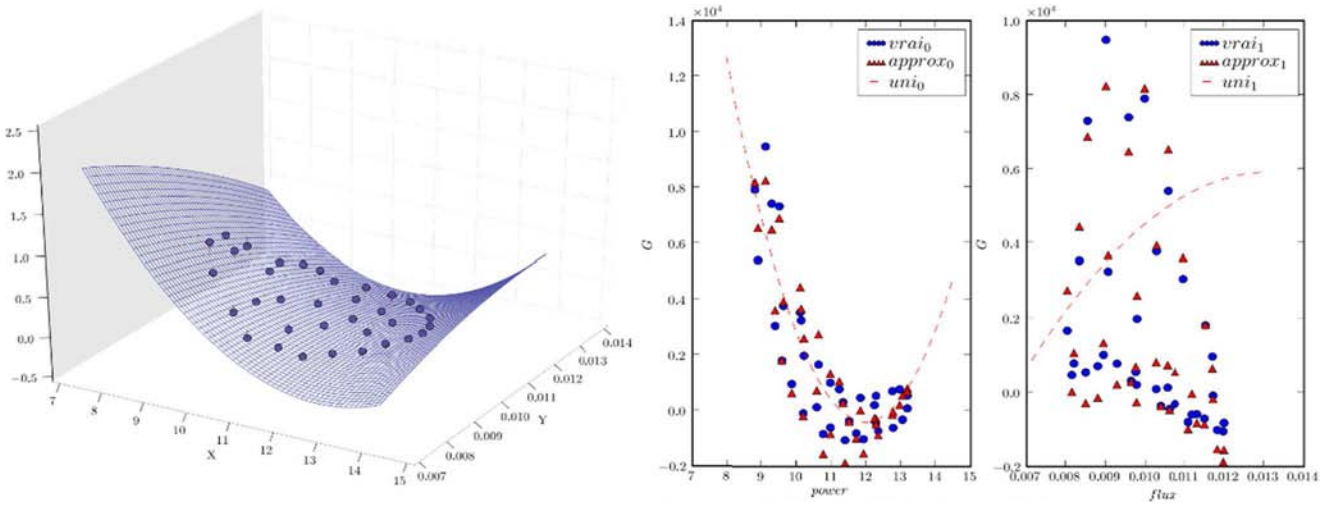


Fig. 11 Response surface for the first iteration

distribution is randomly drawn, respectively. Figure 10c and d show the distributions of regression indices R^2 and $\overline{R^2}$ for the random design of experiments. With extremely good values for both indices, it becomes clear that they cannot be used to assess the quality of the response surface. There are observed, first, a large variability of index β and, second a mean deviating strongly from the real value. Optimizing the design of experiments does not bring much improvement at this level. On the other hand, the value Q^2 (varying between $-\infty$ *to* 0) is very poor for these designs of experiments, which justifies the use of additional experiments for the evaluation of the response surface parameters.

Two new designs of experiments (random and optimized) are then performed, in which a sufficient number of experiments is added to achieve an indicator Q^2>0.90. One can observe that the reliability index dispersion falls, and its mean μ_β=3.23 approaches the real value. A repetition of the procedure is made for the characterization of the response surface around the approximated design point.

The result for the density of the reliability index β is represented in Fig. 10f. One can observe a degradation of the quality index Q^2 for the same number of observations, a mean μ_β=2.7 that deviates from the real value, and a dispersion σ_β higher than for the first iteration. This is because around the design point, the influence of the noise on the limit state function is more important. The design of experiments with closer points then loses it quality. Consequently, the iteration of the response surface is not always necessary. A new rule is defined in the algorithm that ends the process when a loss of quality in the response surface is observed.

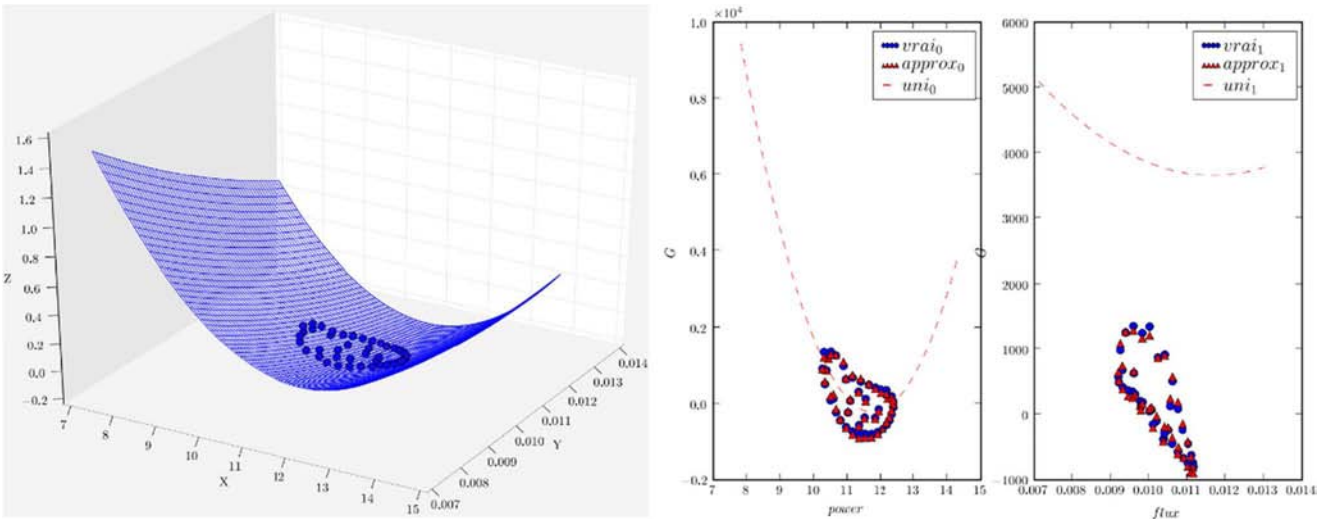


Fig. 12 Response surface for the second iteration

Example 2 – Consider, now, as an input to the model, two basic random variables for the reliability assessment: the coefficient of expansion and the power term of the behaviour law. Both variables are assumed to follow a log-normal distribution: $x_1 = CTE_{S_n/A_g} = \mathscr{L}N\left(2.10^{-5}, 1.10^{-6}\right), x_2 = n = \mathscr{L}N(11, 0.2)$.

The interest of this case lies in the fact that the actual behaviour of the output of the mechanical model is strongly conditioned by the exponential behaviour of variable n. Quadratic response surfaces cannot, then, find a faithful formulation over the entire domain (Fig. 10). Iterating the regression allows approaching the solution and accurately assessing the design point. At the end, the experimental points are properly approached by the quadratic form, and the exponential behaviour is not valid anymore (Fig. 11). Q^2 reaches, then, the value of 0.98. The sensitivity analysis and a SORM calculation works with good accuracy because, in both cases, they are estimated from the location of the design point together with the surrounding points.

The case presented in this paper presents a strong nonlinearity in the number of cycles before failure. This is most likely wrong, given that the regression is only valid in the design point neighbourhood (Figs. 11 and 12). The direct FORM, SORM and response surface methods will struggle to take it into account; only the Monte Carlo approach will be able to highlight it. The time savings cannot be compared because, even in a parallel context, the traditional FORM method requires 24 iterations, or 48 calculation time units (gradients being evaluated at the same time and the evaluation of the timestep for the iHLRF algorithm being also parallelized). The response surface method requires 4 units on a 10-node cluster (20 calls to obtain an acceptable Q^2 coefficient and 2 iterations x 10 calls around the design point). For a number of important variables, with differentiation by a centred finite difference or in higher-level optimization algorithms, the direct FORM method can still be interesting because the load of the cluster is then better optimized when calculating gradients.

The index of reliability and the design point in physical space represented in Table 3 for both methods are almost identical. This demonstrates the ability of the algorithm to find the right design point and properly assess the sensitivities by iterating the calculation loop.

Table 3 Reliability analysis results for both methods

	Direct FORM	Adaptive response surface
β_{HL}	2.73	2.66
P^*	2.13E-0.5	2.16E-0.5
	13.6	13.3

6 Conclusion

Assessing structural reliability in today's industry is widely supported by finite element techniques. The reliability calculations requiring many calls to the FE code, techniques should be implemented to reduce the computation time which may otherwise be prohibitive. Methods based on the use of response surfaces have been introduced in this purpose. The paper presents a method for obtaining a response surface guaranteed by quality indexes in the case of the assessment of an uncertain limit state surface. It relies on a numerical design of experiment capable of optimizing the number of call to the code on the basis of a required quality level. The parallelization of the algorithm of response surface has been investigated in order to reduce the computing time. An illustration of the principles of use of adaptive response surfaces has been proposed on elements of a railway traction system ensuring a function of power switches.

The innovative nature of this work lies in the coupling between the reliability methods and a finite element code for the presentation of a new concept of the so-called adaptive response surface. With the aim of reducing the number of assessment of the limit state function requiring calls to FE code, this method relies on indices of the regression quality in a parallel computing environment. It also allows the refinement of the solution around the design point by iterating the procedure. The prospects for continuation of the work undertaken include analysis of the finite element model to evaluate the accuracy and sensitivity with respect to the mesh, to the law of behaviour as well as to the error on the law integration. In a practical framework, the developments will be conducted for the analysis of power electronics modules in order to study the electric, thermal and mechanical reliability of the modules for which new internal or external uncertainty variables mainly associated with the fabrication process, will have to be identified and quantified.

References

Akay HU, Liu Y, Rassaian M (2003) Simplification of finite element models for thermal fatigue life prediction of pbga packages. J Electron Packag 125:347–354

Alibrandi U (2014) A response surface method for stochastic dynamic analysis. Reliab Eng Syst Saf 126:44–53

Ben Hassen W, Auzanneau F, Incarbone L, Péres F, Tchangani AP (2013a) OMTDR using BER estimation for ambiguities cancellation in ramified networks diagnosis. IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp 414–419

Ben Hassen W, Auzanneau F, Péres F, Tchangani AP (2013b) Diagnosis sensor fusion for wire fault location in CAN bus systems. IEEE SENSORS, pp 1–4

Castellazzi A, Ciappa M (2008) Novel simulation approach for transient analysis and reliable thermal management of power devices. Microelectron Reliab 48(8–9):1500–1504

Chang T-C, Lee C-C, Hsieh C-P, Hung S-C, Cheng R-S (2015) Electrical characteristics and reliability performance of IGBT power device packaging by chip embedding technology. Microelectron Reliab 55(12):2582–2588

Ciappa M (1997) Some reliability aspects of IGBT modules for high-power applications. Thèse de doctorat, Swiss Federal Institute of Technology, Zurich

Darveaux R, Banerji K (1992) Constitutive relations for tin-based solder joints. In IEEE Trans. CHMT, 15(6):1013–1024

Deshpande AM, Subbarayan G (2000) Decomposition techniques for the efficient analysis of area-array packages. J Electron Packag 122(1): 13–19

Engelmaier, W. (1991). Solder Attachment Reliability, Accelerated Testing, and Result Evaluation, chapitre 17, pages 545–587. Van Nostrand Reinhold, New York

Feller L, Hartmann S, Schneider D (2008) Lifetime analysis of solder joints in high power IGBT modules for increasing the reliability for operation at 150 °C. Microelectron Reliab 48(8–9):1161–1166. doi:10.1016/j.microrel.2008.07.019

Gayton N, Bourinet J, Lemaire M (2003) CQ2RS : a new statical approach to response surface method for reliability analysis. Struct Saf 25:99–121

Haukaas T (2003a) Finite element reliability using matlab. http ://www.ce.berkeley.edu/haukaas/FERUM/ferum.html

Jayant Baliga B (2015) Physics, design and applications of the insulated gate bipolar transistor. ISBN: 978-1-4557-3143-5 2015

Jensen HA, Mayorga F, Papadimitriou C (2015) Reliability sensitivity analysis of stochastic finite element models. Comput Methods Appl Mech Eng 296:327–351

Jin, R. (2004). Enhancements of Metamodeling Techniques in Engineering Design. PhD thesis University of Illinois at Chicago

Kanchanomai C, Yamamoto S, Miyashita Y, Muthoh Y, Mcevily AJ (2002) Low cycle fatigue damage analysis under multiaxial random loadings. Int J Fatigue 21:132–140

Lau JH, Pao Y-H (1997) Solder joint reliability of BGA, CSP, flip chip, and fine pitch SMT assemblies. McGraw-Hill Professional

Lhommeau T, Perpiñà X, Martin C, Meuret R, Mermet-Guyennet M, Karama M (2007) Thermal fatigue effects on the temperature distribution inside IGBT modules for zone engine aeronautical applications. Microelectron Reliab 47(9):1779–1783

Lind NC, Hasofer AM (1974) Exact and invariant second-moment code format. J Eng Mech 100(1):111–121

Liu B, Liu D, Tang Y, Chen M (2011) The investigation on the lifetime prediction model of IGBT module. Energy Procedia 12:394–402. doi:10.1016/j.egypro.2011.10.053

Micol A (2007) Approche probabiliste dans la conception des modules de puissance; PhD Thesis, INPT-ENIT

Micol A, Martin C, Dalverny O, Mermet-Guyennet M, Karama M (2008) Probabilistic approaches and reliability design of power modules. Int J Simul Multidiscip Des Optim 2(2):149–156

Micol A, Martin C, Dalverny O, Mermet-Guyennet M, Karama M (2009) Reliability of lead-free solder in power module with stochastic uncertainty. Microelectron Reliab 49(6):631–641

Micol A, Martin C, Mermet-Guyennet M (2011) Viscoplastic behaviour reliability analysis of IGBT module solders. 10th IFAC Workshop on Intelligent Manufacturing Systems (IMS'10), Lisbon, Portugal

Nelhiebel M, Illing R, Detzel T, Wöhlert S, Auer B, Lanzerstorfer S, Ladurner M (2013) Effective and reliable heat management for power devices exposed to cyclic short overload pulses. Microelectron Reliab 53:1745–1749

Pedersen KB, Østergaard LH, Ghimire P, Popok V, Pedersen K (2015) Degradation mapping in high power IGBT modules using four-point probing. Microelectron Reliab 55(8):1196–1204

Pérès F, Grenouilleau J-C (2002) Initial spare parts supply of an orbital system. Aircraft Engineering and Aerospace Technology 74(3)252–262, MCB UP Ltd

Pérès F, Martin C (1999) Design methods applied to the selection of a rapid prototyping resource. 7th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA'99. 1, pp 417–422

Powell, M. J. D. (1994). A direct search optimization method that models the objec- tive and constraint functions by linear interpolation. In Advances in optimization and numerical analysis. S. Gomez and J.P. Hennart, editors

Rackwitz R, Fiessler B (1978) Structural reliability under combined load sequences. Comput Struct 9:489–494

Roussouly N, Petitjean F, Salaun M (2013) A new adaptive response surface method for reliability analysis. Probab Eng Mech 32:103–115

Suhir E (2013) Could electronics reliability be predicted, quantified and assured? Microelectron Reliab 53(7):925–936

Syed, A. (2004). Accumulated creep strain and energy density based thermal fatigue life prediction models for snagcu solder joints. In ECTC 2004 conference

Vallon J (2003) Introduction à l'étude de la fiabilité des cellules de commutation à l'IGBT sous fortes contraintes. PhD thesis, Institut National Polytechnique de Toulouse

Van der Broeck CH, Conrad M, De Doncker RW (2015) A thermal modeling methodology for power semiconductor modules. Microelectron Reliab 55(9–10):1938–1944

Van Driel WD, Zhang GQ, Janssen JHJ, Ernst LJ (2003) Response surface modeling for nonlinear packaging stresses. J Electron Packag 125(4):490–497

Wang GZ, Cheng Z, Becker K, Wilde J (2001) Applying and model to represent the viscoplastic deformation behavior of solder alloys. J Electron Packag 123:247–253

Zou T, Mahadevan S, Mourelatos Z, Meernik P (2002) Reliability analysis of automotive body-door subsystem. Reliab Eng Syst Saf 78(3): 315–324