An efficient adaptive database sampling strategy with applications to eddy current signals

Roberto Miorelli, Xavier Artusi, Christophe Reboud

An efficient adaptive database sampling strategy with applications to eddy current signals

R. Miorelli[a,*], X. Artusi[a], C. Reboud[a]

[a] *Département Imagerie Simulation pour le Contrôle, CEA, LIST, Gif-sur-Yvette 91191, France*

Abstract

Computer simulations are widely used in engineering domains to model complex scenarios and extract meaningful information or improve the understanding of a given problem. Common purposes of simulation studies are inversion, optimization, sensitivity analysis and evaluation of performance. In such contexts, it is often convenient to replace the time consuming forward solver by a metamodel acting as a fast and accurate substitute in a restricted range of input parameters. Focused on applications in the field of Electromagnetic-Non Destructive Testing (E-NDT), this paper proposes an approach to design robust metamodels, based on adaptive databases of simulation results in order to ensure their accuracy. They can then be used as real-time emulators of the physical model and considerably speed up time consuming studies like estimation of probability of detection, defect characterization or sensitivity analysis. The database and metamodel generation problem is first addressed with a meshless approach based on Augmented Radial Basis Function (A-RBF) algorithm. Then, its performance is compared with that of a more standard approach exploiting a n-dimensional Delaunay mesh. Both approaches rely on an adaptive generation technique known in

*Corresponding Author. Tel.: +33169085057 Fax: +33169087597
Email address: roberto.miorelli@cea.fr (R. Miorelli)

the literature as Output Space Filling (OSF). Performance in terms of computational time and results accuracy of both methods are finally evaluated and compared in the case of a specific application: the simulation of Eddy Current Testing (ECT) inspection problems.
Keywords: Database, metamodel, output space filling design, database adaptive sampling, augmented radial basis function, Delaunay mesh, eddy current, non destructive testing.

1. Introduction

In the field of Non Destructive Testing & Evaluation (NDT&E), physical models are commonly used by engineers in order to better understand experimental signals, design components, or evaluate the performance of inspection procedures. In the last two decades, numerical simulation tools have widely spread in the community. As a consequence, new kinds of NDT&E studies, which largely employ numerical simulations have been popularized. Among them, one can cite the Model Assisted Probability Of Detection (MAPOD) [1], Sensitivity Analysis (SA) and defect characterization through parametric inversion. A common characteristic of such studies is the necessity of a large amount of information, implying the computation of many simulated signals (up to several tenths of thousands). Such a large number of simulations makes solution of problems too time consuming when using the models directly. In order to overcome this issue, some research has recently been focused on finding an efficient and general replacement of standard forward solvers [2, 3, 4, 5], consisting in a regression over a database of simulation results built in a restricted range of input parameters. In a first step, also known as off-line phase, this database is adaptively built in order to maximize the fidelity of its associated interpolator (called metamodel), and used

to generate signals in almost real time in the second step (the on-line phase).

1.1. Adaptive database generation using kernel-based methods

Kernel-based database generation has recently been applied with success in the field of NDT. In [6], a Radial Basis Function (RBF) interpolator over an adaptively filled database of simulation results has been proposed. The sampling strategy implemented implicitly relies on a fictitious mesh (even if the RBF interpolator is, in principle, designed to be a meshless approach). To adaptively build the database, both the RBF metamodel and the physical model were evaluated at the center of the edges of the Delaunay mesh connecting the existing points together. Then, new points, for which a significant discrepancy between model and metamodel is observed, are added to the database. In spite of the good performance in terms of robustness and accuracy, the approach badly scaled with respect to the size of the input space. Indeed, the physical solver needed to be called many times all over the mesh to possibly add a small amount of points, thus the database generation could rapidly "explode" in terms of computation time. It is worth mentioning that in [6] an inversion procedure based on particle swarm optimization has been also proposed and tested on a three-dimensional database. An alternative and actually very effective way to tackle the sampling problem was proposed in [7, 8]. In these works, a new sampling strategy, called Output Space Filling (OSF), has been associated to a functional Ordinary Kriging (OK) interpolator. The principle of OSF is to regularly distribute signals in the database, with respect to their variations, which is measured with a distance (or dissimilarity) indicator between them. Sampled points locations in the output space thus tend to be evenly spread with respect to this distance, which definition is a key stone of the method and depends on the nature of the signals at hand.

This OSF-Ordinary Kriging (OSF-OK) scheme has achieved a high interpolation accuracy. Furthermore, good parsimony in terms of number of sampling points needed to build the database, has been observed compared to the previous sampling scheme. Moreover, it was shown in [7] that OK can be used to solve inverse problems, too. Unfortunately, due to the mathematical structure of its kernel, which involves the calculation of a covariance matrix based on a Matérn function [7], this interpolator can be costly to setup from a large amount of samples. For Eddy Current Testing (ECT) applications, in the authors' experience, the generation procedure becomes very difficult when the number of samples exceeds two thousands. In typical ECT problems, limitation makes OK not suitable for generating ECT signal databases when the input space exceeds about six dimensions [9].

1.2. Adaptive database generation using a Delaunay mesh

Beside kernel-based methods, a completely different approach of database generation, using a meshing strategy, has recently been proposed [10, 11]. Two kinds of adaptive sampling strategies through mesh refinement and piecewise constant and piecewise linear interpolation have been proposed, respectively. In [12] an OSF-based n-dimensional Delaunay mesh and a linear interpolator have been employed for database generation, the obtained metamodel being dedicated to parametric inversion based on quadratic programming. Generally speaking, the main drawbacks of the mesh-based approach are related to the fact that a refinement of a mesh in a n-dimensional parameter space can be neither trivial nor very fast to perform. Indeed, database refinement in an input space with more than six dimensions, can easily turn into a very cumbersome and time consuming task when the number of samples increases.

1.3. Paper scope and structure

In this paper, the physical model of interest is used to simulate eddy current testing (ECT) signals, consisting of a set of coil impedance variations or voltage with respect to the probe scan over the inspected material. The stored signal is thus a collection of complex values (up to several thousands when considering 2D maps). Due to the vector or matrix nature of the ECT signals, in the following we explicitly intend sampling strategies able to deal with functional outputs (i.e., vector output). The algorithms presented here apply of course to real valued signals and scalar outputs, too.

The paper is organized as follows. In the first part, the OSF sampling paradigm is jointly applied with an Augmented-RBF (A-RBF) interpolation [13, 14]. Then, a slightly modified OSF sampling technique, compared to [12], has been developed for the Delaunay mesh-based strategy. The modified approach considerably speeds up the database generation process for large input space dimensions without deteriorating, in an appreciable way, the metamodel accuracy. The two proposed solutions have been applied to database generation on realistic test cases. The first case is associated to the nuclear domain with steam generator tube inspection and the second one to the aeronautic domain with the inspection of planar multilayered structures. In order to show the robustness of the proposed sampling strategy, six- and eight- dimensions databases have been generated for nuclear-related and aeronautic-related test case, respectively. In order to assess the quality of the generated database and thus the associated metamodel results, a Cross Validation (CV) procedure has been carried out. Furthermore, through error analysis of CV results we show how one can employ those data in order to retrieve "for free" meaningful meta-information on the metamodel prediction accuracy. The last part of this paper discusses the obtained results with a

highlight on respective advantages and drawbacks of both methods, introducing additional developments that can be envisaged in perspective of this work.

2. Database generation through OSF sampling

Let us describe the physical model of interest by a deterministic forward operator $\mathcal{F} : \mathbb{R}^D \mapsto \mathbb{C}^M$, the vector $\mathbf{x} = [x_1, x_2, ..., x_D]$ being the set of D input parameters such that $\mathbf{x} \in \mathbb{R}^D$. By applying $\mathcal{F}$ to the input vector parameter, we obtain the corresponding output vector $\mathbf{y} = \mathcal{F}\{\mathbf{x}\}$, where $\mathbf{y} = [y_1, y_2, ..., y_M]$ such that, generally speaking, $\mathbf{y} \in \mathbb{C}^M$ and M represents output cardinality. Then we define the set of the input space parameters as $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ..., \mathbf{x}_N\}^\intercal \setminus \mathbf{x}_i \in \mathbb{R}^{1\times D}$ the associated output space as $\mathbb{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_i ..., \mathbf{y}_N\}^\intercal \backslash \mathbf{y}_i = \mathcal{F}\{\mathbf{x}_i\} \in \mathbb{C}^{1\times M}$ with N is the final number of samples considered. Therefore, a database made by N input/output couples is straightforwardly defined as $\mathbb{D} = [(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_N, \mathbf{y}_N)]^\intercal$ where the superscript $^\intercal$ stands for the transpose operator.

As previously stated, the OSF criterium aims to fill a database $\mathbb{D}$ such that the samples are spread evenly in the output space $\mathbb{Y}$ with respect to a certain metric, in this work the $\mathcal{L}^2$-norm accordingly to [7]. Thus, the distance between any point in the output space $\mathcal{F}\{\mathbf{x}\}$ and the i-th vector $\mathbf{y}_i$ within the database is given as

$$Q_i(\mathbf{x}) = \|\mathcal{F}\{\mathbf{x}\} - \mathbf{y}_i\|_2, \qquad \mathbf{x} \in \mathbb{X}, \tag{1}$$

where we notice that $Q_i(\mathbf{x})$ is function of the input space and is computed directly in the output space. In order to fill the database through OSF strategy, at every iteration, we look for the best candidate input vector among a representative pool of randomly chosen trials (or targets) $\mathbf{x}_t \in \mathbb{X}_T$ such

that $\mathbb{X}_T = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ..., \mathbf{x}_T\} \setminus \mathbf{x}_t \in \mathbb{X}$, with T is the number of candidate samples considered. Therefore, from (1) we compute all the distances $Q_i(\mathbf{x}_t)$ between the trial candidates and all the points already present within the database. To avoid calling directly the time consuming forward solver $\mathcal{F}$ for all the T candidates, one can replace it with an auxiliary metamodel $\mathcal{M} : \mathbb{R}^D \mapsto \mathbb{C}^M$. We define $\hat{\mathbb{Y}}\{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, ..., \hat{\mathbf{y}}_t ..., \hat{\mathbf{y}}_T\} \setminus \hat{\mathbf{y}}_t = \mathcal{M}\{\mathbf{x}_t\} \in \mathbb{C}^{1\times M}$. That is, $\mathcal{M}$ behaves as a suitable interpolating function calculated from the actual database $\mathbb{D}$. In this way (1) turns into

$$\hat{Q}_i(\mathbf{x}_t) = \|\mathcal{M}\{\mathbf{x}_t\} - \mathbf{y}_i\|_2, \qquad \mathbf{x}_t \in \mathbb{X}, \tag{2}$$

where $\hat{Q}_i(\mathbf{x}_t)$ represents the *estimated* output distance between the t-th candidate and the i-th sample within the database. The point $\mathbf{x}_t^*$, selected to be added to the database, is the one that maximizes the minimum distance $\mathbf{x}_t^* = \underset{\mathbf{x_t}\in\mathbb{X}_t}{argmax}\left\{\underset{i\in 1,2,...,N}{min}\hat{Q}_i(\mathbf{x}_t)\right\}$. In other words, the next model evaluation is always carried out at the location of the input space where its immediate neighborhood will be as far as possible –it fills the "largest hole". The generation of the set of targets samples $\mathbb{X}_T$ relies on a Latin Hyper-cube Sampling (LHS) design. Other kind of pseudo-random or deterministic sequences can be employed straightforwardly. Once the OSF stopping criterion is met, the database filling procedure ends. Further insight to OSF sampling algorithm can be found in the pseudo-code listed in Algorithm 1. For deeper details one can refer to [7]. In this work we have compared the meshless interpolator based on A-RBF interpolation algorithm and the mesh-based one relying on a Delaunay mesh with a multilinear interpolator.

2.1. OSF scheme with A-RBF interpolation

In equation (2) we have defined $\hat{Q}_i(\mathbf{x}_t)$ without mentioning explicitly the interpolation function associated to the metamodel $\mathcal{M}$. In this subsection,

OSF adaptive sampling strategy **1**

begin

▷ *initialize the database with N samples*

$\mathbb{D} = [(\mathbf{x}_1, \mathbf{y}_1 = \mathcal{F}\{\mathbf{y}_1\}), (\mathbf{x}_2, \mathbf{y}_2 = \mathcal{F}\{\mathbf{y}_2\}), ..., (\mathbf{x}_N, \mathbf{y}_N = \mathcal{F}\{\mathbf{y}_N\})]$

$Iter = 0$, $Error_{Iter=0} = \inf$

▷ *start the adaptive loop*

while $Iter <= Iter_{max}$ or $Error_{Iter} > Error_{max}$

▷ *generate set of T candidate parameters*

▷ *with Latin Hypercube Sampling*

$\mathbf{x}_T = LHS(T, D)$

▷ *evaluate metamodel on $\mathbf{x}_t$*

$\hat{\mathbf{y}}_t = \mathcal{M}\{\mathbf{x}_t\}$ with $t = 1, ..., T$

▷ *calculate pair distance between points in output space*

$Q_i(\mathbf{x}) = \|\mathcal{M}\{\mathbf{x}_t\} - \mathbf{y}_i\|_2$ with $i = 1, ..., N$

▷ *find new point to add to database*

$$\mathbf{x}_t^* = \underset{\mathbf{x_t} \in \mathbb{X}_t}{argmax} \left\{ \min_{i \in 1,2,...,N} \hat{Q}_i(\mathbf{x}_t) \right\}$$

▷ *run forward solver on new candidate sample*

$\mathcal{F}\{\mathbf{x}_t^*\}$

▷ *updating database by adding new sample ($N = N + 1$)*

$\mathbb{D} = [(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_N, \mathbf{y}_N), (\mathbf{x}_t^*, \mathcal{F}\{\mathbf{x}_t^*\})]$

▷ *calculate $NL2$ mean error trough cross validation*

$Error_{Iter} = \frac{1}{N} \sum_{i=1}^{N} NL2E(\mathbf{x}_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{\|\mathbf{y}_i - \mathcal{M}\{\mathbf{x}_i\}\|_2}{\|\mathbf{y}_i\|_2}$

▷ *update counter*

$Iter = Iter + 1$

end while

end

we deal with the definition of $\mathcal{M}$ by briefly introducing the structure of the A-RBF interpolator employed in this work. Let us define the A-RBF interpolation function as [13, 14]

$$\hat{\mathbf{y}}\left(\mathbf{x}_t\right) = \mathcal{M}\left\{\mathbf{x}_t\right\} = \sum_{i=1}^{N} \lambda_i \phi\left(\left\|\mathbf{x}_t - \mathbf{x}_i\right\|_2\right) + \sum_{k=1}^{K} \mu_k p_k\left(\mathbf{x}_t\right) \tag{3}$$

where λ_i is a vector representing the coefficients of RBF weights, $\phi\left(\left\|\mathbf{x}_t - \mathbf{x}_i\right\|_2\right)$ is the kernel function. The latter sum over k is done on an additional D-variate polynomial term p_k with degree less than or equal to j such that $K = \begin{pmatrix} j+D \\ j \end{pmatrix}$ where μ_k are coefficient to be determined. The last sum of the right hand side of (3) represents the "augmented" part which distinguishes the A-RBF from the standard RBF. The polynomial part of degree j is employed to avoid oscillations between sampling points and it ensures that the A-RBF reproduces a j-degree polynomial behavior between the different nodes (more detail can be found in [13, 14]). The additional K degrees of freedom are fixed by imposing the the following auxiliary condition

$$\sum_{i=1}^{N} \lambda_i p_k\left(\mathbf{x}_i\right) = 0, \qquad k = 1, 2, ..., K. \tag{4}$$

Furthermore, in order to satisfy the interpolation requirements, e.g., the interpolation passes trough the support nodes $\mathbf{x}_i$ and that in equation (6) the rank$(P) = K$, the following constraints are imposed

$$\sum_{k=1}^{K} \mu_k p_k\left(\mathbf{x}_i\right) = 0, \qquad i = 1, 2, ..., N. \tag{5}$$

The coefficients λ_i and μ_k are determined from the interpolation equations (3) and (5) associated to the set of N computed samples, by solving the

following system of equations

$$\begin{bmatrix} \Phi & | & \mathbf{P} \\ - & + & - \\ \mathbf{P}^{\intercal} & | & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ - \\ \mu \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ - \\ 0 \end{bmatrix}, \tag{6}$$

where Φ is a $N \times N$ is a positive defined symmetric matrix also known as kernel matrix. $\mathbf{P}$ is a $N \times K$ matrix (and $\mathbf{P}^{\intercal}$ is its transpose), λ is a $N \times 1$ vector, μ is a $K \times 1$ vector and $\mathbf{y}$ is a $N \times M$ matrix. The kernel function employed in this work is known in the literature as thin plate spline for which $\Phi_{ti} = \phi\left(\|\mathbf{x}_t - \mathbf{x}_i\|_2\right) = \phi\left(\mathbf{r}_{ti}\right) = \mathbf{r}_{ti}^2 \log\left(\mathbf{r}_{ti}\right)$. This kind of kernel has been chosen since the Φ-matrix can be calculated without any time-consuming estimation of shape parameters, which is almost mandatory for other kinds of kernels [15, 16]. In this study, first degree augmented polynomial were used, therefore $\mathbf{P}$ has size the $N \times D + 1$ and forms

$$\mathbf{P} = \begin{bmatrix} 1 & \mathbf{x}_1 \\ 1 & \mathbf{x}_2 \\ \vdots & \vdots \\ 1 & \mathbf{x}_N \end{bmatrix}.$$

One of the main differences between OK (and kriging in general) and A-RBF is that, with the latter, the weights λ and the coefficients μ are retrieved once from matrix system (6) and then are employed in the interpolation stage through a simple matrix product. In OK, the calculation of the weights (called also kriging coefficients) depends on the correlation between the already embedded database samples and the target one. Therefore, OK prediction involves a first stage in which the correlation matrix is evaluated for all the T target samples. Then the linear system of equation is solved in order to obtain the suitable weights. Finally the weights are multiplied by the

associated database output values in order to obtain the desired prediction. All these steps turn the interpolation process into a more time consuming task compared to A-RBF.

2.2. OSF scheme with local Delaunay refinement

In this section, we analyze a modified OSF scheme, with respect to the one proposed by [12], for generating an adaptive database through Delaunay mesh using the Matlab function `delaunayn` based on Qhull algorithm [17]. The metamodel results are obtained by evaluating it on a set of candidate points $\mathbf{x}_t$ with $t = 1, ..., T$. The linear interpolation over the Delaunay mesh made by S simplexes is defined as

$$\hat{\mathbf{y}}\left(\mathbf{x}_t\right) = \mathcal{M}\left\{\mathbf{x}_t\right\} = \sum_{s=1}^{S} b_s\left(\mathbf{x}_t\right) \cdot \left(\mathbf{B}_s \mathbf{x}_t + \mathbf{c}_s\right), \tag{7}$$

where $b_s\left(\mathbf{x}_t\right)$ is equal to 1 if the t-th sample $\mathbf{x}_t$ belongs the s-th simplex and it is 0 otherwise. Coefficients $\mathbf{B}_s$ and $\mathbf{c}_s$ are calculated such that $\hat{\mathbf{y}}\left(\mathbf{x}_i\right) = \mathbf{y}\left(\mathbf{x}_i\right)$ with $i = 1, 2, ..., N$ belongs to the index of the sampled points. In [12] equation (7) must be solved for the whole number of T randomly chosen trials candidates samples for which, each time, a searching procedures among the S simplexes must be performed to retrieve the suitable simplexes subset (i.e., if the samples belong or not to a simplex). This searching process may turn to be very time consuming for databases having high cardinality and large number of samples, indeed its complexity increases as $\mathcal{O}\left\{N^{D/2}\right\}$. In order to mitigate these drawbacks we propose an alternative OSF scheme in which the set of candidate points $\mathbf{x}_t$ belongs to the barycentre of each D-dimension polyhedral compounding the mesh. This choice allows us to obtain directly the hyper-space locations of trial points using a lookup table. Additionally, through the coordinates of the barycentre (e.g., the trial points locations)

and the associated polyhedral vertexes, we are able to establish the set of hyper-spheres (i.e., the D-dimension extension of the circumcircles in the 2D case) including the barycentre coordinates. This knowledge enables us to automatically detect the minimum set of simplexes that should be refined, then classical local refinement strategies can be applied [18]. That is, the local generation of the Delaunay mesh avoids the complete (re-)generation of the Delaunay mesh through more time consuming Matlab function `delaunayn`. Moreover, by choosing $\mathbf{x}_t$ as belonging to the barycentre of each polyhedral we skip the time consuming phase linked to the searching phase done through the Qhull-based Matlab function `tsearchn`. It is worth mentioning that the choice of generating 6-dimensions and 8-dimensions databases has been done also to assess the performance of a Delaunay mesh-based approach. Indeed, for the latter database we almost reach the algorithmic limit of Delaunay triangulations, which is heuristically estimated around 8-dimensions accordingly to [17].

3. Results of ECT database generation through OSF scheme

In order to propose test cases close to realistic problems for which experimental benchmark or experimental measurements are available, we have chosen to address two different ECT inspection problems already presented in the literature. The first generated database deals with a steam generator tube inspection for which a public benchmark has been proposed by Iowa State University [19]. The latter database has been generated for a test case involving a planar multilayered structure for which experiments have been performed at University of West Macedonia by a quite prolific team in terms of ECT benchmark problems [20, 21, 22]. In the next section we apply both sampling strategies proposed in Section 2 for generating the aforementioned

databases. Then the associated metamodels have been compared in terms of computational time for interpolating signals and accuracy with respect to the forward model.In the analyzed test cases, we are interested in obtaining ECT signals issued from a time-harmonic excitation applied to the driven coil. The quantity sensed by the pick-up coil, which represents the ECT signal, is in our case the coil impedance variation, which is a complex valued.

3.1. First database: steam generator tube test case

The first test case deals with a steam generator like inspection configuration (see Fig. 1). Starting from the set of nominal values specified in [19], we have built the database by choosing 6 main variables of interest from the inspection point of view. These parameters are the crack dimensions in terms of length (l) and angular extensions (i.e., width w), the probe lift-off (lo), its tilt angles with respect to the tube axis (θ_A) and the tube orthoradial direction (θ_C). In order to address the whole benchmark cases [19], the complete set of benchmark frequencies (f) have been considered as database parameter too. This increases quite a lot the database complexity. In fact, within the frequency range considered, the ECT signals behave very differently in terms of amplitude and phase for a given set of the remaining database input parameters. The forward model is the one of the CIVA software [23], based on a dedicated integral equation approach [24]. A rectangular (91×201 coil positions) map surrounding the crack zone has been simulated, therefore each sample within the database contains 18290 complex-valued measurements points. In Fig. 2 we show an example of ECT signal associated to this test case. The amplitude map of the coil impedance variation is shown in Fig. 2(a) as well as two signals extractions. These extractions are shown in terms of real and imaginary part in the complex plane in Fig. 2(b).

The database initialization has been done by considering, within the in-

put parameters validity range shown in Tab. 1, a scheme based on a coarse (i.e., 1215 samples) full-factorial grid to which we add a set of 485 samples through Latin hypercube sampling design. This allows to fill the input space homogeneously with a better parsimony than a standard full factorial scheme. Then, by applying the OSF adaptive sampling, 278 samples have been added iteratively. In order to stop the sampling process, we have defined a suitable error to asset the quality of the metamodel prediction. The chosen error is based on the following Normalized $\mathcal{L}^2$-norm Error (NL2E)

$$NL2E\left(\mathbf{x}_i\right) = \frac{\left\|\mathcal{F}\left\{\mathbf{x}_i\right\} - \mathcal{M}\left\{\mathbf{x}_i\right\}\right\|_2}{\left\|\mathcal{F}\left\{\mathbf{x}_i\right\}\right\|_2}, \qquad \mathbf{x}_i \in \mathbb{X}, \tag{8}$$

where $\mathbf{x}_i$ correspond to the inputs combination under check, which does not belong to the database used to define $\mathcal{M}$. The denominator of NL2E in eq. (8), contains the $\mathcal{L}^2$-norm of the difference between forward solver results and metamodel prediction, respectively for a given input. The denominator contains the $\mathcal{L}^2$-norm of the ECT signal obtained through the solver evaluated on the input $\mathbf{x}_i$. In the following, we refer to the mean NL2N as the average value of eq. (8) over the set of tested samples evaluated in the cross validation procedure. To check the whole set of database samples, we have applied a 10-K fold cross validation process on the whole set of database samples. The average NL2E distribution obtained via this validation process is shown Fig. 3 in case of A-RBF interpolation and in Fig. 4 in case of the linear interpolation based on Delaunay mesh. In these pictures, error values are plotted up to 50% for readability. The number of outliers lying outside of the interval, in terms of percentage, can be retrieved via the cumulative red curve values.

By looking at the cross validation errors (e.g., see Fig. 3 and Fig. 4), we can readily notice that the error distribution associated to the two meta-

Id.	Parameter		Validity Range
1	Coil Lift-off (lo)	[mm]	$[0.005; 0.35]$
2	Coil Tilt Axial (θ_A)	[deg]	$[-5; 5]$
3	Coil Tilt Circ. (θ_C)	[deg]	$[-5; 5]$
4	Working Freq. (f)	[kHz]	$[10; 250]$
5	Crack Length (l)	[mm]	$[8; 15]$
6	Crack Ang. Ext. (w)	[deg]	$[0.11; 1]$

Table 1: Steam generator tube case, database parameters with associated validity range.

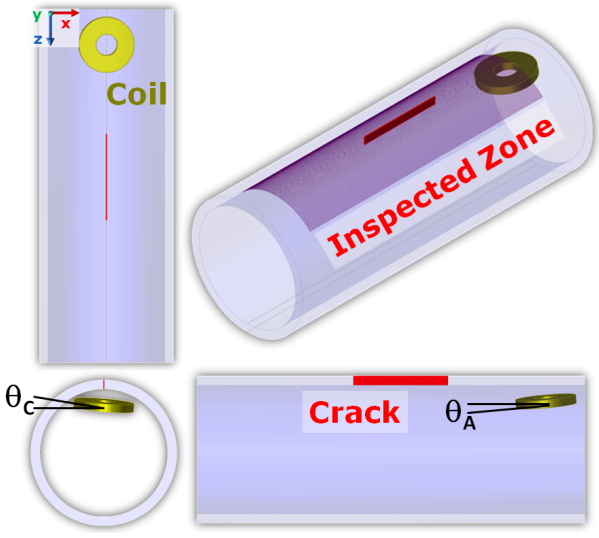
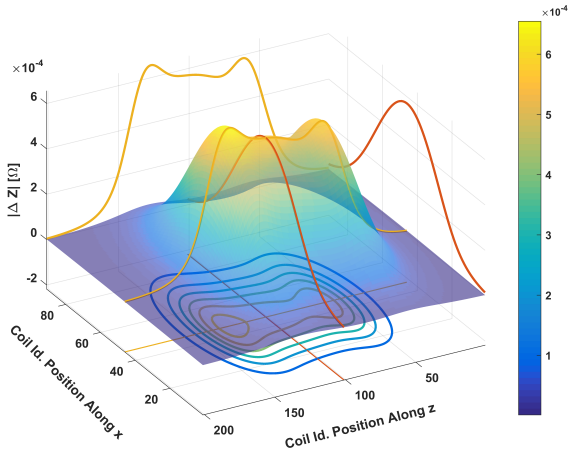
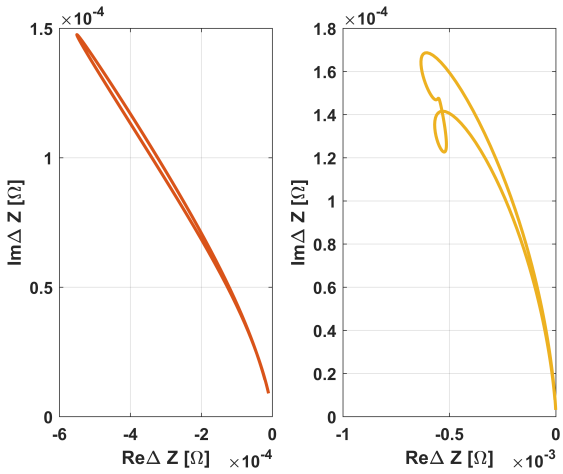


Figure 1: Different views of the tube inspection configuration [19].



(a)



(b)

Figure 2: Example of ECT signal associated to the tube case inspection. In (a), the impedance variation is shown in terms of amplitude together with two orthogonal signal extractions for a coil inspection passing across and along the crack. In (b), both extraction are shown in the complex plane.

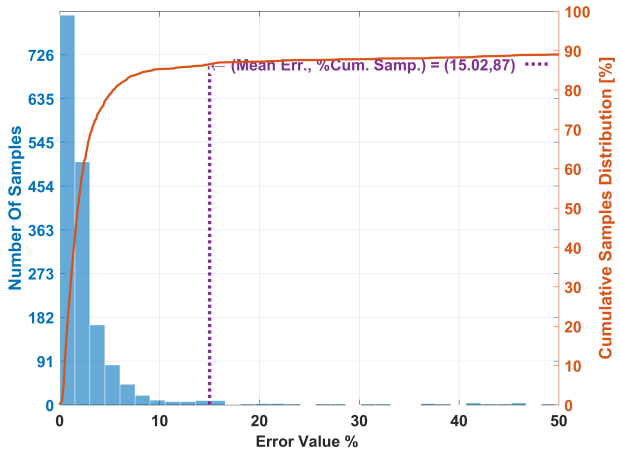


Figure 3: Results of metamodel check accuracy, realization done through 10-K fold cross validation based on the first database. The surrogate model for the tube test case is obtained via A-RBF interpolation. NL2E vs. number of samples distribution are shown. Error distribution through histograms is displayed (left hand y-axis), and the solid line represents the cumulative number of samples with respect to the error value (right hand y-axis).

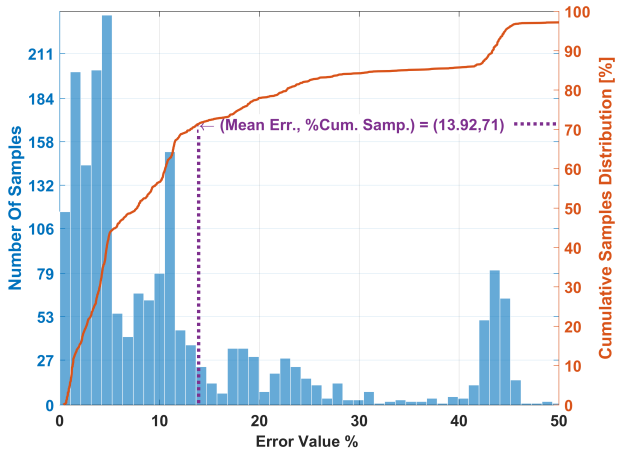


Figure 4: Results of metamodel check accuracy, realization done through 10-K fold cross validation based on the first database. The surrogate model for the tube test case is obtained via Delaunay linear interpolation. NL2E vs. number of samples distribution are shown. Error distribution through histograms is displayed (left hand y-axis), and the solid line represents the cumulative number of samples with respect to the error value (right hand y-axis).

models behaves quite differently. Indeed, A-RBF distribution sports a longer tail compared to the Delaunay mesh-based approach which, conversely, has a lower dumping rate with respect to A-RBF that shows a very high rate samples concentrated at the lower error values. In order to better determine how the error distribution behaves in the output space, in Fig. 5 and Fig. 6 we have displayed scattered plot of comparisons between true values (i.e., simulations) vs. metamodel results for A-RBF and linear interpolation, respectively. Each dot on these two scattered plots corresponds to $\mathcal{L}^2$-norm values calculated from CIVA and metamodel signal maps. We can notice that, for A-RBF metamodel, the higher the values of errors (see Fig. 3), the smaller the values of the $\mathcal{L}^2$-norm, this means that A-RBF interpolation is not able to describe accurately very small ECT signals. On the other hand, the mesh-based metamodel outputs have some bias in results which turns into a more spread scattered plot, as we can notice in Fig. 6. The set of points showing the bias corresponds to the errors for which the histogram in Fig. 4 presents local changes in bars distribution with respect to a decreasing behaviour.

3.2. Second database: planar layered structure

The second database has been generated by considering the three layers planar structure inspected by a coil working in absolute mode (see Fig. 7) for which information on experimental set-up and experimental data are available in [25]. The database has also been built by using CIVA software [23], which provides very good results in this case. For simulating the inspection procedure, a rectangular map of 71×71 points surrounding the cracks zone has been considered at the simulations stage. Therefore, for each sample within the database contains 5040 complex-valued measurements points. In Fig. 8 we show an example of of ECT signal for the tube test case. The

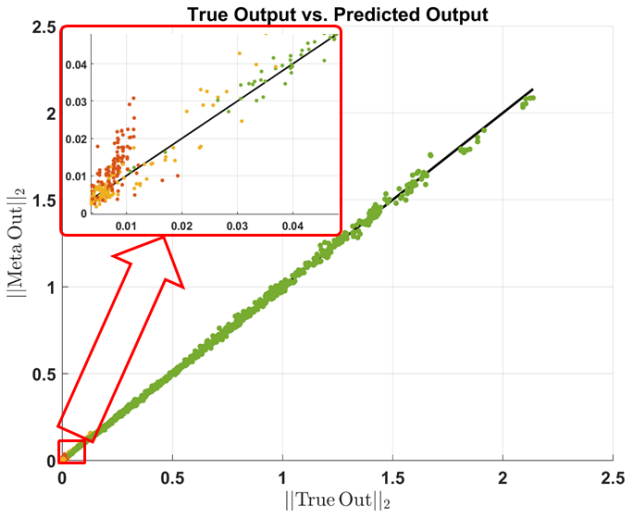


Figure 5: Metamodel results analysis applied on the first database. 10-K fold cross validation results for the $\mathcal{L}^2$-norm values obtained from the complete map signal: true model (i.e., CIVA) vs. A-RBF metamodel.

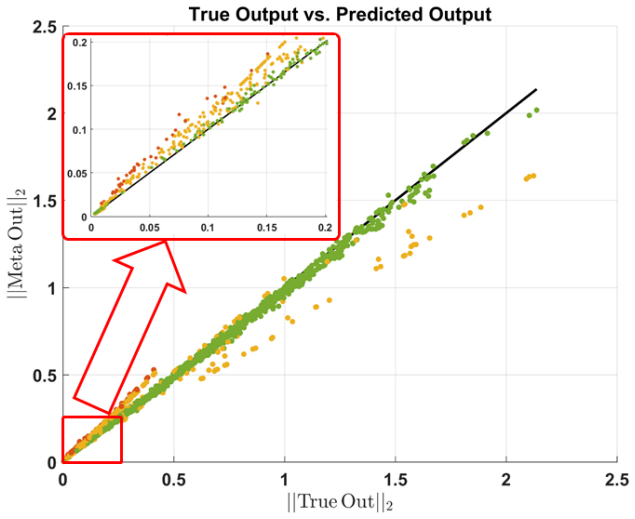


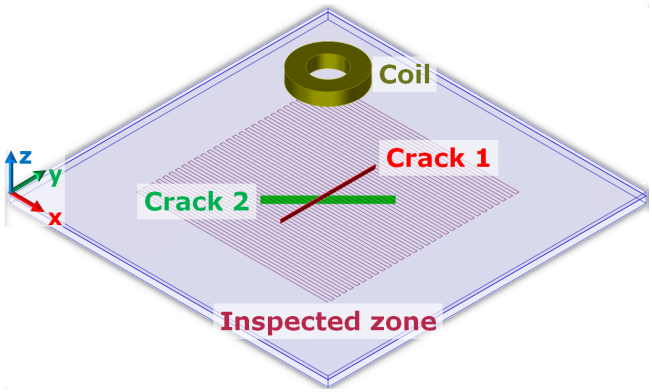
Figure 6: Metamodel results analysis applied on the first database. 10-K fold cross validation results for the $\mathcal{L}^2$-norm values obtained from the complete map signal: true model (i.e., CIVA) vs. linear metamodel.

Id.	Parameter		Validity Range
1	Coil Lift-off (lo)	[mm]	$[0.025; 0.25]$
2	Crack 1 Length (l_{C1})	[mm]	$[25; 33]$
3	Crack 1 Width (w_{C1})	[mm]	$[0.05; 0.5]$
4	Crack 2 Width (w_{C2})	[mm]	$[0.05; 0.5]$
5	Crack 2 Length (l_{C2})	[mm]	$[25; 33]$
6	Crack 2 Skew Angle (ϕ)	[deg]	$[-15; 105]$
7	Coil Tilt (θ)	[deg]	$[0.11; 1]$
8	Dielectric Thickness (d)	[mm]	$[0.05; 0.5]$

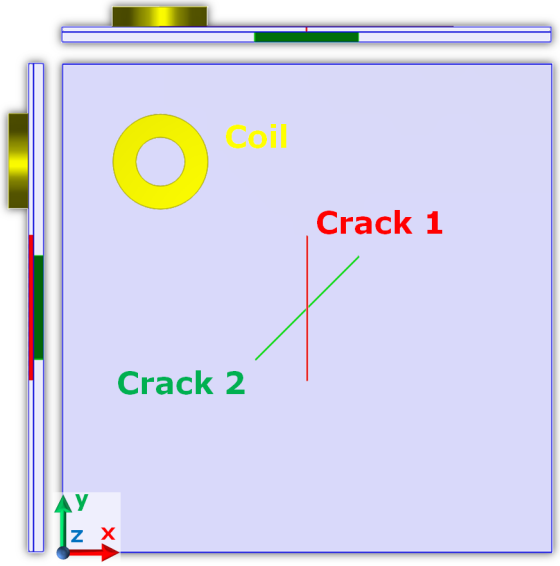
Table 2: Planar layered structure database parameters with the associated validity range.

amplitude the coil impedance variation and two signal extractions are shown in the complex plane in Fig. 8 (a) and (b), respectively. It is worth noticing that these signals behave differently from the previous test case, due to the different geometry and inspection scenario considered (multiple cracks in layered medium).

This second database has been built with eight input parameters, for which the validity range is given in Tab. 2. In particular, starting from a coarse full factorial grid initialization made of 1944 (i.e., $3^5 \times 2^3$) samples, a set of 800 Latin hypercube samples has been taken as initialization, before applying the adaptive sampling based on OSF with A-RBF interpolator. The sampling process has been stopped once the 10-K fold CV has achieved a mean NL2E error (see eq. (8)) less than 7%, which led to a final database size equal to 3544 samples. Results obtained through the mesh-based approach via linear interpolation has not been presented due to the computational effort needed to generate the suitable database. More details on this point and related aspects are discussed in Section 3.3.2.

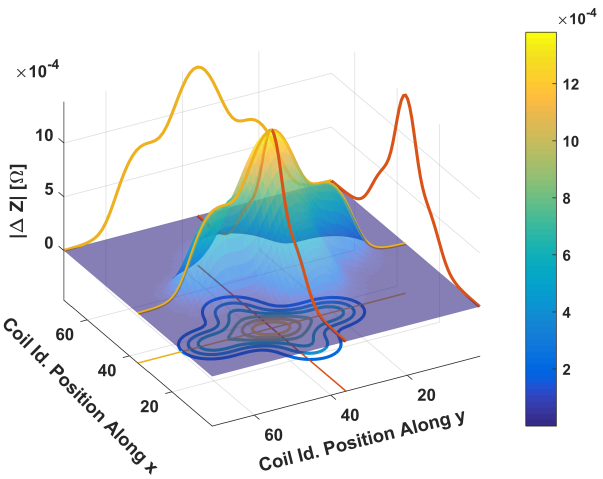


(a)

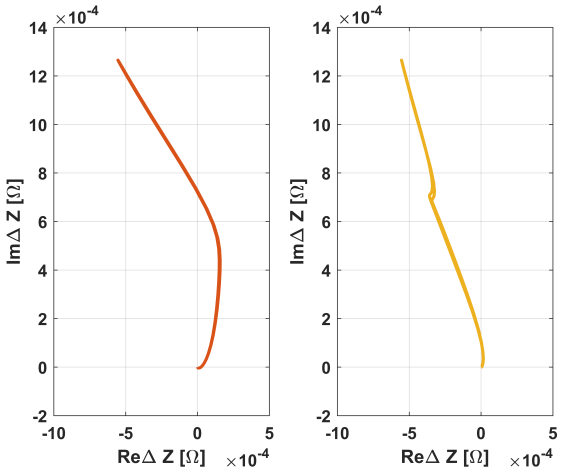


(b)

Figure 7: Planar multilayer structure test case. (a) 3-D view. (b) Top and side views.



(a)



(b)

Figure 8: Example of ECT signal associated to the planar multilayer structure inspection. In (a), the impedance variation is shown in terms of amplitude together with two orthogonal signal extractions for a coil inspection passing across and along the crack. In (b), the complex valued impedance variation signal associated to the extraction are shown in the complex plane, respectively.

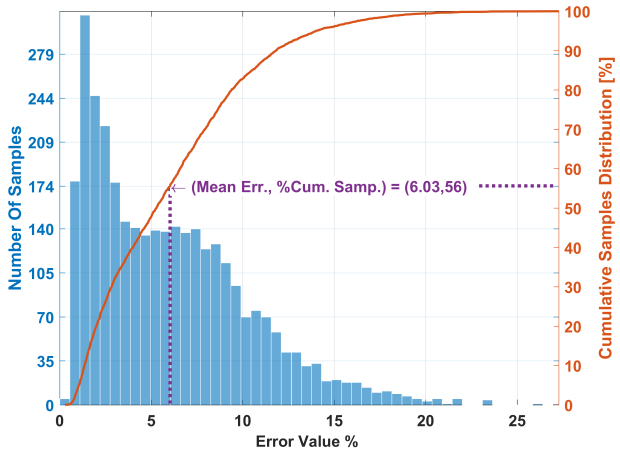


Figure 9: Results of metamodel check accuracy, realization done through 10-K fold cross validation based on the second database. The surrogate model for the planar test case is obtained via A-RBF interpolation. NL2E vs. number of samples distribution are shown. Error distribution through histograms is displayed (left hand y-axis) and the solid line represents the cumulative number of samples with respect to the error value (right hand y-axis).

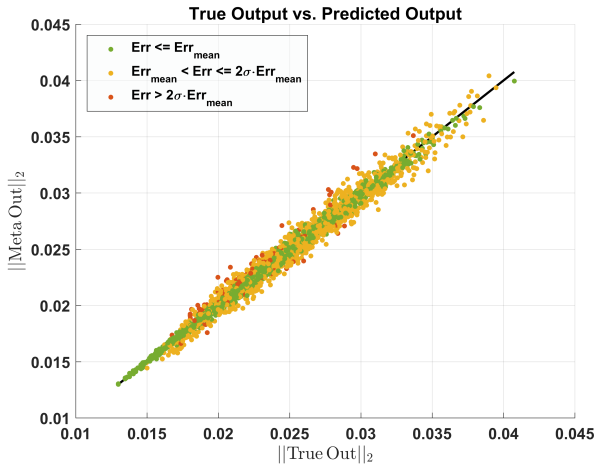


Figure 10: Metamodel results analysis applied on the second database. 10-K cross validation results for the $\mathcal{L}^2$-norm values obtained from the complete map signal: true model (i.e., CIVA) vs. A-RBF metamodel.

The scattered plot in Fig. 10 shows comparisons in terms of $\mathcal{L}^2$-norm obtained through 10-K fold CV between true and predicted values for the A-RBF metamodel. We can notice that results are homogeneously spread around 45° dark line which identifies the full accordance between "real" and predicted values. This behavior can be seen as an alternative view of envelop of the error distribution shown in Fig. 9, where points far from the solid black line correspond to the higher errors. The metamodel accuracy is very satisfactory in this case.

3.3. Remarks and comments on the results

3.3.1. Meta-information embedded within the databases

In Section 3.1 and Section 3.2, the evaluation of database generation results obtained through A-RBF and Delaunay mesh-based approaches has been introduced. In both test cases, 10-K fold CV has been employed to estimate the NL2E error of each metamodel, leading to results shown in Fig. 3 and Fig. 9 for A-RBF, and in Fig. 4 for the Delaunay-based linear interpolation. Those plots show that the error distribution has a different shape with each technique, as indicated by standard variation values collected in Tab. 3. Moreover, we can see on this table that the A-RBF results are globally as accurate as the mesh-based ones, if the mean NL2E is considered. In addition, in the case of A-RBF the error is confined to a region corresponding to small ECT signals in terms of magnitude (e.g., small flaws), which is very convenient for NDT applications. As a side note, despite the lesser number of input parameters, the tube case was more difficult to handle for both metamodels. This is due to the frequency parameter, which spans over a wide range of values, making the impedance variation changing consistently for a given set of the other parameters. This means, somehow, that the function to interpolate is more complex for this test case. Finally, for the planar case,

Test Case Type		Tube	Planar
A-RBF	$\langle$NL2E$\rangle$	$\sim 14.8\%$	$\sim 6\%$
	std. dev.	$\sim 37.9\%$	$\sim 4.3\%$
Linear	$\langle$NL2E$\rangle$	$\sim 14.7\%$	*n.c.*
	std. dev.	$\sim 15.9\%$	*n.c.*

Table 3: Comparisons of mean NL2E and standard deviation calculated via 10-K fold cross validation of results shown in Fig. 3 and Fig. 9 for A-RBF and in Fig. 4 for Delaunay mesh-based interpolation. The values are averaged over ten different trials.

the error calculation in Tab. 3 is not available for linear interpolation, since the CV process has been too heavy to be handled.

The CV procedure gives access to some additional meaningful information. Indeed, for each sample belonging to the histograms shown in Fig. 3 and Fig. 9, an error map can be created to predict the accuracy (i.e., the error) of the interpolation in a circumscribed zone of the database (i.e., the test set location). In practice, a database of error values can be automatically created with the aim to built an associated metamodel (on the errors values) to be exploited at the prediction stage. Therefore, two databases one from the training phase, i.e., $\mathbb{D} = [(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_N, \mathbf{y}_N)]$, and one from error analysis through CV, i.e., $\mathbb{D}_\varepsilon = [(\mathbf{x}_1, \varepsilon_1), (\mathbf{x}_2, \varepsilon_2), ..., (\mathbf{x}_N, \varepsilon_N)]$, can be employed to predict the metamodel output at the unknown input location $\mathbf{x}_t$ as $\hat{\mathbf{y}}(\mathbf{x}_t) = \mathcal{M}\{\mathbf{x}_t\} + \mathcal{M}_\varepsilon\{\mathbf{x}_t\}$. In order to visualize the errors of the database $\mathbb{D}$ in a higher-dimensional space, we shall borrow a common tool employed in statistic and bioengineering science, which is called parallel coordinates plots [26]. In Fig. 11, we show the parallel coordinates plot obtained through the error analysis done within the cross validation process based on A-RBF metamodel when the planar test case is considered. The solid thick lines, having colors going from yellow to red, associate the errors ranging from 15% to

maximum value $\sim 27.3\%$ with the database input parameters (i.e., the index "Id." in Tab. 2) and corresponding values in the normalized parameter space given in the vertical position. The background thin lines, coloured from light yellow to green, represent samples having errors values less than 15%. Taking into consideration Tab. 2, Fig. 11 shows that errors tend to be concentrated for the highest values of the length associated to crack 2 (i.e., Id. 5), the lowest values of the crack angle for deepest crack (i.e., Id. 6) and the extreme values of the dielectric thickness (i.e., Id. 8). Thus, by employing the parallel coordinates plot shown in Fig. 11, we can empirically evaluate the quality of metamodel predictions provided a set of input parameters through the colorbar representing the error magnitude. Moreover, if a particular pattern is contained in the parallel coordinates plot (as it is for parameters Ids. 5, 6, and 8 in Fig. 11), then it means that in some zones of the input parameters space, large signal variations occur and consequently bad approximations are likely there. In order to better interpret the nature of a particular pattern, we can plot an alternative parallel coordinates plot gathering the information of particular set of samples within the database. In Fig. 12 we show the impact of the adaptive OSF strategy in the input space. We can notice that the OSF strategy have the tendency to add samples inducing a particular pattern in the input parameter space. In particular, lower values of probe lift-off (Id. 1), longer crack 2 length (Id. 5) and smaller dielectric thickness (Id. 8) have been more sampled than other locations for these parameters. From the physical point of view, the aforementioned parameter combinations are the ones providing the highest variation of the ECT signal, which indicates a certain relevance of the proposed sampling strategy. Along the other dimensions i.e., crack 1 length (Id. 2), crack 1 width (Id. 3) and crack 2 width (Id. 4) samples are more concentrated at the extremities of the domain, which is ex-

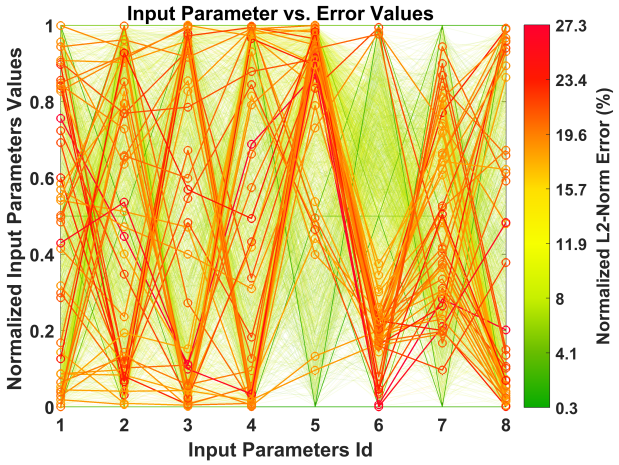


Figure 11: Parallel coordinates plot obtained from 10-K fold cross validation results associated to the A-RBF metamodel for the planar test case (see Fig. 9). Each line represents a database point in a 8-dimensional space and the color associated highlight the prediction error calculated through cross validation.

pected. In fact, we notice that the smaller values of crack 1 width (Id. 3) and crack 2 width (Id. 4), which provide smaller ECT signals are "compensated" by larger values of the crack 1 length (Id. 3) (which provides the stronger ECT signals). The distribution of samples for the angular position of crack 2 (Id. 6) is less easy to interpret. However, a secondary preferential pattern is shown when the two cracks tends to be aligned i.e., the upper part of Id. 6 range which. The coil tilt seems to show a secondary tendency to focus on a particular interval. As last remark, if we superpose the plot shown in Fig. 11 with the one in Fig. 12, we notice that the database area refined by OSF adaptive sampling adds samples nearby zones where the error is the greater one (except for the crack 2 angular position for which this is only partially true).

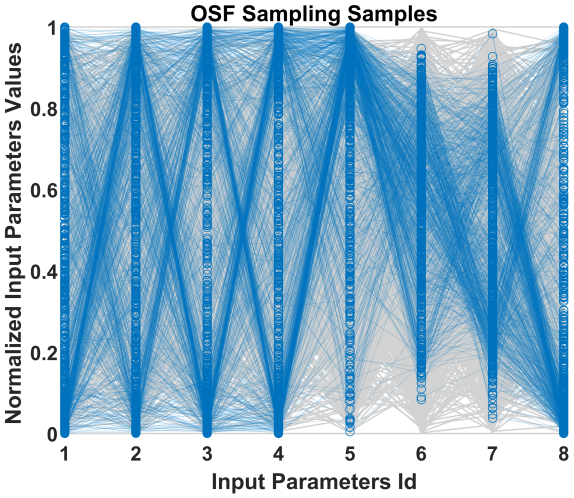


Figure 12: Parallel coordinates plot showing the impact of the OSF adaptive sampling strategy in the input parameter space for the planar test case (see Fig. 9). Light gray lines on the background represent the initialization samples and the foreground blue lines with marker 'o' represent the added samples through OSF.

3.3.2. Computational time associated to the metamodel evaluation

Tab. 4 gives the computational time for A-RBF and Delaunay-based linear interpolations for both cases treated. We can readily notice that the first one largely outperforms the second one. The main reason of such a gap in computational time is due to the fact that the generation of a Delaunay mesh must be done at each iteration of the CV process. Furthermore, an exhaustive research on the whole set of simplexes of the Delaunay mesh is needed for the each test sample. Moreover, still in Tab. 4, we can notice that computational time for planar case through linear interpolation is not available. Indeed, we were not able to finalize the CV process on a desktop Dell Precision T7810 equipped with an Intel Xeon E5 – 2630 having 32 GB of RAM due to the free space RAM problem, whereas all the other results in here presented have been generated on a laptop Dell Precision M3800 equipped with Intel i7-4270 HQ with 16 GB RAM. The 10-K fold CV process requires

Test Case Type	Tube	Planar
Interp. Time A-RBF [sec]	~ 80	~ 44
Interp. Time Linear [sec]	~ 1896	*n.c.*
Num. Metamodel Evaluations	1915	3288

Table 4: Comparisons of interpolation time for the A-RBF and the Delaunay mesh-based interpolation needed to perform cross validation results shown in Fig. 3 and Fig. 9 and in Fig. 4, respectively.

to randomly split the entire database into ten parts and then it uses nine parts as training set (i.e., as database) and one part as test set (i.e., set of samples to be evaluated by the metamodel), then the CV procedure stops when all the different portions have been employed as test set. Since both training and test samples can be considered as a quasi-random distributed, at each CV stage an 8-dimensions Delaunay mesh over about 3000 samples must be generated through the Matlab function `delaunay`. Successively, the Matlab function `tsearchn` is employed to look for simplexes involved in the interpolation of the about 328 test set samples located in an 8-dimensional space. Since the complexity of the Qhull algorithm scales as $\mathcal{O}\left\{N^{D/2}\right\}$, the interpolation stage may be very time and resource consuming when the database has high cardinality and large number of samples as it is for the proposed test cases. This justifies the developed approach for generating the database with linear interpolation. As a matter of fact, by imposing the location of candidates points to be added in the adaptive loop at each barycentre location, we completely avoid the time consuming research of the candidates points within the Delaunay mesh.

It may be noted that, due to the high cardinality of the input space (D) for both databases studied, a large number of samples is already reached even if a very coarse initialization is done. This aspect clearly sets a limit

for other kind of mesh less based interpolation like OK [7, 8] which becomes cumbersome or even impossible to use in these cases when the number of input dimension exceeds about 6 [9]. Furthermore, the combination of A-RBF and OSF sampling criteria allows to handle problems better than the mesh-based counterpart from both the accuracy and the computational point of view. Moreover, the very large gap in terms of computational time between the Delaunay mesh-based and the A-RBF, makes the latter approach a more suitable metamodel candidate to be employed for intensive simulation campaigns.

4. Conclusions and perspectives

An OSF-based sampling strategy has been presented to adaptively generate a database and its associated metamodel, with a focus on applications to ECT simulation. A mesh-less A-RBF interpolator and a Delaunay mesh-based interpolator have been implemented and compared, respectively. The performance of both metamodels were analysed in terms of computational time and accuracy through a cross validation process. This analysis shows that A-RBF clearly outperforms the Delaunay based approach in both aspects. Furthermore, the mesh-based approach is not able to provide results in a reasonable time when the input dimensions increases over a number of about eight. Thus, it looses its usefulness for time consuming tasks like, for instance, MAPOD, SA and stochastic inversion algorithms.

From a general point of view, the OSF sampling scheme has confirmed to be a valuable tool for database generation of ECT signal focused on parsimony. Concerning the interpolation phase, other techniques like the RBF-QR one [27] and compactly supported RBF [28, 29] can be employed to increase the A-RBF performances for even higher database dimensions. Moreover,

regression scheme based on Learning-By-Example (LBE) techniques [30] can also be used to provide an alternative approach with *a-priori* promising results. Concerning LBE techniques, we highlight that once a database and the associated metamodel are available, both regression in terms of surrogate forward and inversion [30] models are available at the same time. Moreover, through the use of the state-of-the-art machine learning algorithms [31], we can use metamodel results in order retrieve more information contents associated to a database. To investigate this particular aspect, feature extraction techniques, like for instance principal component analysis, partial least squares and their kernel counterparts, could lead to a more robust database generation strategy. Besides, the cross validation procedure has given us access to a certain level of meta-information. In particular, it enables to distinguish zones in which the error overcomes a certain threshold. One axis of next developments could be to extract such critical areas and run independently an additional OSF refinement on each of them, in order to improve accuracy at a low computational cost.

5. Acknowledgments

This work was done in the framework ANR ByPASS project. Authors would like to thank E. Demaldent for the very fruitful discussions around Delaunay mesh refinement scheme and the problems related to database generation procedures.

References

[1] P. Calmon, F. Jenson, and C. Reboud, Simulated probability of detection maps in case of non-monotonic EC signal response, AIP Conference Proceedings (2015)1933-1939.

[2] A. Forrester, A. Sobester, A. Keane, Engineering Design Via Surrogate Modelling: A Practical Guide, Wiley, New York, 2008.

[3] K. Crombecq, E. Laermans, T. Dhaene, Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling. European Journal of Operational Research 214(2011) 683-696.

[4] E.J. Chen, M. Li, Design of experiments for interpolation-based metamodels, Simulation Modelling Practice and Theory 44(2014) 14-25.

[5] L. Van Gelder, P. Das, H. Janssen, S. Roels, Comparative study of metamodelling techniques in building energy simulation: Guidelines for practitioners. Simulation Modelling Practice and Theory 49(2014) 245-257.

[6] R. Douvenot, M. Lambert, D. Lesselier, Adaptive metamodels for crack characterization in eddy-current testing. IEEE Transactions on Magnetics 47(2011) 746-755.

[7] S. Bilicz, M. Lambert, S. Gyimòthy, Kriging-based generation of optimal databases as forward and inverse surrogate models. Inverse Problems 26(2010) 074012.

[8] S. Bilicz, E. Vazquez, S. Gyimòthy, J. Pàvò, M. Lambert, Kriging for eddy-current testing problems. IEEE Transactions on Magnetics 46(2010) 4582-4590.

[9] S. Bilicz, Sparse grid surrogate models for electromagnetic problems with many parameter. IEEE Transactions on Magnetics 52(2016) 1-4.

[10] S. Gyimòthy, S. Kiss, J. Pàvò, Adaptive sampling technique based on moving meshes for building data-equidistant inversion databases for NDT,

International Journal Applied Electromagnetic and Mechanics 30(2009) 309-319.

[11] S. Gyimòthy, Optimal sampling for fast eddy current testing inversion by utilising sensitivity data, ET Science, Measurement & Technology 9(2015) 235-240.

[12] S. Bilicz, Inversion of eddy-current testing signals using a fast interpolation over an optimal defect-database, K. Capova, L. Udpa, L. Janousek, B.P.C Rao (Eds.) Electromagnetic Non-Destructive Evaluation, Studies in Applied Electromagnetics and Mechanics, IOS Press, Amsterdam, 2013, pp. 56-57.

[13] T. Poggio and F. Girosi, Networks for approximation and learning. Proceeding IEEE 9(1990) 1481-1497.

[14] G.B. Wright, Radial basis function interpolation: numerical and analytical developments. University of Colorado PhD. Dissertation 2003.

[15] S. Rippa, An algorithm for selecting a good value for the parameter c in radial basis function interpolation. Advances in Computational Mathematics 11(1999) 193-210.

[16] M. Mongillo, Choosing basis functions and shape parameters for radial basis function methods, SIAM Undergraduate Research Online 4(2011) 190-209.

[17] http://www.qhull.org/html, 2016 (accessed 11.08.16).

[18] S. Rebay, Efficient unstructured mesh Generation by means of Delaunay triangulation and Bowyer-Watson algorithm. Journal of Computational Physics 106(1993) 125-138.

[19] http://www.wfndec.org/benchmarkproblems_files/2012%20EC%20Benchmark%20Announcement.png, 2016 (accessed 07.07.16).

[20] http://www.wfndec.org/benchmarkproblems_files/2013%20EC%20Benchmark%20Announcement.png, 2016 (accessed 07.07.16).

[21] http://www.wfndec.org/benchmarkproblems_files/2014%20EC%20Benchmark%20Combined.png, 2016 (accessed 07.07.16).

[22] http://www.wfndec.org/benchmarkproblems_files/2015%20EC%20Benchmark_with_Data.zip, 2016 (accessed 07.07.16).

[23] www-civa.cea.fr, 2016 (accessed 14.07.16).

[24] J.R. Bowler, T. Theodoulidis, H. Xie, Y. Ji, Evaluation of eddy-current probe signals due to cracks in fastener holes. IEEE Transactions on Magnetics 48(2012) 1159-1170.

[25] R. Miorelli, C. Reboud, T. Theodoulidis, D. Lesselier, Efficient modeling of ECT signals for realistic cracks in layered half-space. IEEE Transactions on Magnetics 49(2013) 2986-2992.

[26] B. Vidakovic, Statistics for Bioengineering Sciences, Springer, New York, 2011.

[27] B. Fornberg, E. Larsson, N. Flyer, Stable computation with Gaussian radial basis function. SAIM Journal Of Scientific Computing 33(2011) 869-892.

[28] H. Wendland, Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Advances in Computational Mathematics 4(1995) 389-396.

[29] Z. Wu, Compactly supported positive definite radial functions. Advances in Computational Mathematics 4(1995) 283-292.

[30] S. Salucci, N. Anselmi, G. Oliveri, P. Calmon, R. Miorelli, C. Reboud, A. Massa, Real-time NDT-NDE through an innovative adaptive partial least squares SVR inversion approach. IEEE Transactions on Geoscience and Remote Sensing 54(2016) 6818-6832.

[31] S. Theodoridis, Machine Learning, a Bayesian and Optimization Perspective, Elsevier Academic Press, London, 2015.