Arch. Mech., 50, 5, pp. 829–847, Warszawa 1998
FIFTY YEARS OF THE ARCHIVES OF MECHANICS

On efficiency of identification of a stochastic crack propagation model based on Virkler experimental data

Z. A. KOTULSKI (WARSZAWA)

IN THE PAPER we concentrate on one aspect of the experimental design: how the information coming from an experiment can be utilised for identification of a specific mathematical model. To express the consistency of the data and the model we need some quality measure, allowing to transform our intuition to numbers. As the mathematical tool we propose a version of the statistical procedure of cross-validation of the data. Then we verify the efficiency of the suggested method on the example of the Virkler experimental data of stochastic crack growth and the mathematical model of Paris-Erdogan of the fatigue crack growth.

1. Introduction

EXPERIMENTAL DATA constitute a basis of the mathematical modelling of physical phenomena. Trying to identify the model's parameters we always ask the question if the data are sufficiently reliable for the applied mathematical procedure. Development of mathematical statistics achieved in recent years made it possible to perform methodologically consistent reasoning to decide whether the obtained experimental results are useful for the proposed model and inversely – whether the model is adequate for the experimental data.

The purpose of the paper is to propose a method of verification of the quality of experimental data coming from some physical phenomenon for identification of a certain mathematical model of this phenomenon. (The same purpose can be written in an inverse way: what is the quality of a certain mathematical model for description of a physical phenomenon generating the observed set of numerical data). After general remarks on collecting the empirical data, we concentrate on a particular model of stochastic crack growth. We make an attempt to verify if the Virkler experimental curves of crack growth can be used for identification of the Paris-Erdogan model of the stochastic crack propagation [10]. The method applied for this purpose is the cross-validation method of verification of predictability of the measured data, widely applied in mathematical statistics (see [1, 5, 11, 12]). At the beginning we present the general (non-linear) formulation of the cross-validation technique. Next we formulate the problem in a linear case and present the formulae for estimation of the linear model parameters when some measurements are missing. Finally we apply the proposed procedure to verification of the Virkler data being the source of knowledge for the simplified Paris-Erdogan model of the stochastic crack growth.

2. Experiment's design and reliability of experimental data

Researchers using experimental data for verification of the mathematical models of physical phenomena have always a dilemma: to make their own experiment or to apply experimental data available in the literature. In both cases they encounter several methodological and technical problems.

Constructing our own experiment, we can do this according to all the rules known as the design of experiment in a way optimal for the specific mathematical model considered [6]. To plan the experiment, one should:

- select the model variables that must be identified;
- select the set of treatments (different factors whose effects are being compared) effecting on the measured quantities;
- specify the experimental material to which the treatments are to be applied;
- construct or select the rules according to which the measured data are connected with the model parameters;
- manipulate the treatments (increase the number of samples, modify the range of controlled experiment parameters, etc.) in such a way that finally, the identified model is possibly complete.

We realise that, in spite of the fact that there is a temptation to manipulate the results of the experiment to improve the quality of identification and validation of the mathematical model (interesting remarks on possible tricks and methods of detecting such manipulations can be found in [9]), one can also really modify the experiment to improve its results. However, sometimes the objective reasons (high cost of experiment, difficulties in keeping constant experiment's conditions, unexpected noises during measurements, etc.) make that the collected data are not satisfactory and one feels to be obliged to verify their validity.

Applying in the modelling procedure the experimental data taken from literature, researchers meet quite different problems. First of all, they never know all the conditions of the experiment. However, even if the description of the experiment itself and of the presented data is sufficient for the modelling purpose, they reach a fundamental barrier: the number of data samples is fixed and cannot be increased by continuation of the experiment. Then they should always answer questions like: Is the set of the experimental data sufficiently large? What would be the effect of estimation if we had more data from the experiment? In other words, one must answer the question if the available experimental data set is sufficiently representative for identification of the proposed mathematical model.

The heuristic idea of verification of experimental data as the basis of identification of the selected mathematical model (the estimation of its parameters) can be formulated in a mathematical way. An example of such a procedure is presented in the following sections.

3. Cross-validation method and estimation

The cross-validation is a method of verifying the consistency of experimental data. In this method we choose two different subsamples from the data sample. One subsample is applied for estimation of the system parameters, the other is used as a reference set to control the quality of estimation. This procedure lets us to test two facts: the integrity of the experimental data (the data sample is in some sense homogeneous if both subsamples of it give similar estimation results), and correctness of the estimation procedure (the algorithm gives similar results for two different subsamples of data taken from the same population).

The standard cross-validation procedure can be modified for any particular problem and any expected purpose of it. Now we present a version of this method useful for verification of the measurements obtained from an experiment.

Consider the following two-dimensional time series:

$$(y_i, x_i), \qquad i = 1, 2, ..., n, \tag{3.1}$$

where the elements of the sequence represent, respectively: x_i – the observed data points, y_i – the values of the process being estimated.

Assume that we know some number of the data pairs $(y_i, x_i), i = 1, 2, ..., n$; we call them the observation history S. Assume also that for the given observation history we can construct the estimator $\hat{y}(x, \boldsymbol{\alpha}, S)$ of the random variable y based on the observation x (the value of the process corresponding to the observation x). In this estimator, the parameter $\boldsymbol{\alpha} \in \mathbf{A}$ ($\boldsymbol{\alpha}$ is some scalar, vector or matrix parameter taking its values from a certain set of parameters $\mathbf{A}$) describes the dependence of the values of the process y_i on the data points x_i, for $i = 1, 2, ..., n$, and it depends on the history S. Parameter $\boldsymbol{\alpha}$ should be also estimated during (or before) the estimation of y. Using the constructed estimator we make an attempt to verify the quality of experimental data using the following cross-validation type procedure.

Consider n observation data points. Assume that a subsample of $n - 1$ data points is used for the estimation of the parameter $\boldsymbol{\alpha}$. We estimate this parameter n times, every time omitting another point. We are interested, how much the omitted data points influence the quality of estimation of $\boldsymbol{\alpha}$ and, consequently, of the process y. To answer this question we define the following scheme of reasoning.

The cross-validation algorithm

I. Estimate the parameter using n-1 samples, minimising the following functional:

$$L(\boldsymbol{\alpha}) = \frac{1}{n-1} \sum_{j=1,2,...i-1,i+1,...,n} L\left[y_j, \hat{y}(x_j, \boldsymbol{\alpha}, S_{/i})\right], \tag{3.2}$$

where $L[\ ,\ ]$ is some loss function and $S_{/i}$ is the observation history of $n-1$ pairs, where the pair (y_i, x_i) is omitted.

II. Apply the procedure of point **I** n times for $i = 1, 2, ..., n$. For each step, fix the estimated value of the parameter $\boldsymbol{\alpha}$ as:

$$\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}(S_{/i}), \quad i = 1, 2, ..., n. \tag{3.3}$$

III. Estimate the states of the observed process y according to the assumed estimation formula, where the parameter is taken as $\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}(S_{/i})$, that is calculate the values $\hat{y}(x_i, \tilde{\boldsymbol{\alpha}}(S_{/i}), S_{/i}), i = 1, 2, ..., n$, minimising the expression:

$$C(S) = \frac{1}{n}\sum_{i=1}^{n} L\left[y_i, \hat{y}(x_i, \tilde{\boldsymbol{\alpha}}(S_{/i}), S_{/i})\right]. \tag{3.4}$$

The value of $C(S)$ calculated in (3.4) for the obtained values of the estimators gives us the quality measure of the estimation procedure.

IV. Estimate the reference values of the process using all the history S. We obtain them by minimising the following functional:

$$C_{\text{ref}}(S) = \frac{1}{n}\sum_{i=1}^{n} L\Big[y_i, \hat{y}(x_i, \boldsymbol{\alpha}(S), S)\Big]. \tag{3.5}$$

Let us remark that in some cases the procedure (3.5) using the complete history S, can give the exact estimated values of the process y, that is $\hat{y}(x_i, \boldsymbol{\alpha}(S), S) = y_i$ and, consequently, $C_{\text{ref}}(S) = 0$. However, for some specific estimators this can not be satisfied, and then we should compare the measures (3.4) and (3.5).

The cross-validation procedure enables us to verify the integrity of the experimental data. It detects, how much information about a single measurement is contained in the rest of the measurements of the observation history. If in the data population there are some outstanding results, they will contribute a significant income to the quality measure (3.4). When the observation history contains a lot of such data points, the value of $C(S)$ becomes much greater than $C_{\text{ref}}(S)$ and we can expect that any increase of the number of data points in the identification procedure can effect in a significant change of the model parameters being estimated.

Let us remark that the procedure of cross-validation is performed for a finite number of data points n. The number n growing to infinity in the validation procedure does not guarantee the convergence of the quality measure $C(S)$.

In the above procedure we have assumed as a reference set, the one-point subsamples. In general one can do this by estimating the model parameter $\boldsymbol{\alpha} \in \mathbf{A}$

and omitting several data points, and then in the verification step using the entire experiment history S. In Sec. 8 we apply such a method at a practical example.

4. Linear estimation for non-complete set of experimental data

In this section we consider the known linear estimation procedure. It proves to be very useful for the cross-validation technique in the case when the process is linearly dependent on the model parameters.

Assume that we have the following set of observations:

$$x_i, \qquad i = 1, 2, ..., n. \tag{4.1}$$

The process to estimate is denoted by:

$$y_i(\boldsymbol{\alpha}), \qquad i = 1, 2, ..., n, \tag{4.2}$$

where $\boldsymbol{\alpha}$ is the (vector) parameter to be fixed during the estimation procedure.

Since the model is assumed to be linear, the process y can be represented as:

$$y_i = \sum_{j=1}^{p} A_{ij}, \alpha_j, \qquad i = 1, 2, ..., n. \tag{4.3}$$

The values of the observations x and the process y are connected by the following observation equation:

$$x_i = y_i + e_i, \qquad i = 1, 2, ..., n, \tag{4.4}$$

or

$$x_i = \sum_{j=1}^{p} A_{ij}\alpha_j + e_i, \qquad i = 1, 2, ..., n, \tag{4.5}$$

where $A_{ij}, i = 1, 2, ..., n, j = 1, ..., p$ are the elements of the system matrix, and $e_i, i = 1, 2, ..., n$ are the elements of the random disturbance (noise) vector.

The formulation of the estimation problem

We assume that our observation process (set of n observations) can be written down in the following matrix form [6]:

$$\mathbf{x} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{e}, \tag{4.6}$$

where

$$\mathbf{x} = (x_1, x_2, ..., x_n)^T \tag{4.7}$$

is the observation vector,

$$\mathbf{A}=\begin{bmatrix} A_{11} & \ldots & A_{1p} \\ . & . & . \\ . & . & . \\ . & . & . \\ A_{n1} & \ldots & A_{np} \end{bmatrix} \tag{4.8}$$

is the system matrix,

$$\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_p)^T \tag{4.9}$$

is the vector of parameters to estimate,

$$\mathbf{e}=(e_1,\ldots,e_n) \tag{4.10}$$

is the noise (random disturbance or error) vector.

For the efficiency of the model it is assumed that

- A_{ij}, the elements of the system matrix, are some known constants
- x_i, the elements of the observation vector, are normally distributed;
- x_i are independent;
- all the variables x_i have identical variance σ^2.

From the above conditions we can deduce that the elements e_i of the noise vector are Gaussian, independent random variables (we assume: with a zero mean) and with identical variance σ^2.

To complete the vector formulation of the problem we rewrite equation (4.3) in the form

$$\mathbf{y}=\mathbf{A}\boldsymbol{\alpha}. \tag{4.11}$$

Then the estimated value the process is

$$\hat{\mathbf{y}}=\mathbf{A}\hat{\boldsymbol{\alpha}}, \tag{4.12}$$

where $\hat{\boldsymbol{\alpha}}$ is the estimated value of the control parameter $\boldsymbol{\alpha}$.

If the rank of the coefficient (system) matrix $\mathbf{A}$ is p, then the matrix $\mathbf{A}^T\mathbf{A}$ is non-singular and the mean-square linear estimator $\hat{\boldsymbol{\alpha}}$ can be expressed as:

$$\hat{\boldsymbol{\alpha}}=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}. \tag{4.13}$$

Having introduced the basic definitions and facts, we are ready to present the linear version of the scheme of cross-validation analogous to the one presented in the previous section. However, in the linear case we assume the reference subsample as a certain k-element subset of the observation history.

Consider the observations $x_1, x_2, \ldots, x_n$. Assume that the observations x_1, $x_2, \ldots, x_{n-k}$ are used for the estimation of the model parameter $\boldsymbol{\alpha}$, and that

$x_{n-k+1}, ..., x_n$ are omitted in this procedure. Then the matrices and vectors in the state equation (4.6) can be reduced to the following form:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix}, \qquad \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}, \tag{4.14}$$

where

$$\mathbf{x}_1 = (x_1, ..., x_{n-k})^T, \tag{4.15}$$

$$\mathbf{x}_2 = (x_{n-k+1}, ..., x_n)^T. \tag{4.16}$$

The other matrices and vectors are uniquely defined by this division of the observation vector.

By assumption (last k observations are missing) we find the mean-square estimator of the parameter from the following state equation:

$$\mathbf{x}_1 = \mathrm{A}_1 \boldsymbol{\alpha} + \mathbf{e}_1, \tag{4.17}$$

that is $\boldsymbol{\alpha}$ is the solution of the following normal equation:

$$\mathbf{A}_1^T \mathbf{A}_1 \hat{\boldsymbol{\alpha}} = \mathbf{A}_1^T \mathbf{x}_1. \tag{4.18}$$

If $\hat{\boldsymbol{\alpha}}$ is the calculated value of the estimator, then we assume

$$\mathbf{x}_2 = \mathbf{A}_2 \tilde{\boldsymbol{\alpha}}, \tag{4.19}$$

as a substitute for the missing observations. Since the normal equation for the complete system is

$$\mathbf{A}_1^T \mathbf{A}_1 \boldsymbol{\alpha} + \mathbf{A}_2^T \mathbf{A}_2 \boldsymbol{\alpha} = \mathbf{A}_1^T \mathbf{x}_1 + \mathbf{A}_2^T \mathbf{x}_2, \tag{4.20}$$

we assume the observed process in the form

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 = \mathbf{A}_2 \tilde{\boldsymbol{\alpha}} \end{bmatrix}, \tag{4.21}$$

and now $\tilde{\boldsymbol{\alpha}}$ is also the solution of the normal equation.

Let us remark that the quality measure used in calculation of $\tilde{\boldsymbol{\alpha}}$ is:

$$C(S) = \frac{1}{n-k} \sum_{i=1}^{n-k} (x_i - A_{i1}\alpha_1 - ... - A_{ip}\alpha_p)^2. \tag{4.22}$$

It is seen that the above formulae (after the appropriate permutation of the variables) can be used for calculations in the cross-validation method presented in Sec. 3 in the linear case.

Let us remark that the procedure of linear estimation of parameters is (under quite general assumptions) asymptotically convergent, that is, if in (4.13) we take into account a sufficiently great number of observations, we obtain as a result the almost exact value of the expectation of the parameter α. However, in our considerations we deal with a finite number of observations and, moreover, apply this estimator at the algorithm of cross-validation which is not convergent itself (see previous Sec. 3). Therefore the cross-validation procedure gives us only qualitative information about the experimental data.

5. Mathematical model of crack growth

In the literature, various models of stochastic crack growth are used [10]. For the purpose of presentation of the cross-validation method we adopt one of the classical models. Consider the following randomised Paris-Erdogan equation for the fatigue crack growth under homogeneous cyclic stressing [2, 3]:

$$\Delta a = XC(\Delta K)^m, \tag{5.1}$$

with

$$\Delta K = \Delta\sigma F\left(\frac{a}{b}\right)\sqrt{\pi a}, \tag{5.2}$$

$$F\left(\frac{a}{b}\right) = \frac{1}{\sqrt{\cos\pi\frac{a}{b}}}, \qquad \text{for } \frac{a}{b} < 0.7, \tag{5.3}$$

where: a is the crack length, b is the specimen width, Δa is the increment of crack length caused by a single stress cycle, ΔK is the range of the stress intensity at the crack tip, C, m are constants depending on the specimen material, $\Delta\sigma$ is the stress range, X is a random variable changing independently from one crack increment to another, and satisfying the following conditions:

$$E\{X\} = 1, \qquad E\left\{(X-1)^2\right\} = \delta. \tag{5.4}$$

The process of the stochastic crack growth modelled by the discrete randomised Paris-Erdogan equation (5.1)–(5.3) can be equivalently described by the following continuous stochastic differential equation [2, 3]:

$$\frac{da}{\left(F\left(\frac{a}{b}\right)\sqrt{\pi a}\right)^m} = C(\Delta\sigma)^m(1+\xi(t))dt. \tag{5.5}$$

Equation (5.5) has been obtained from (5.1) under the following essential assumption on the random variable X:

$$X = 1 + \xi(t), \tag{5.6}$$

where $\xi(t)$ is a white noise with a zero mean and the intensity δ. The time parameter t is considered to be the number of cycles of the external excitation of the material sample.

Equation (5.5) can be integrated at time intervals $[N_i, N_{i+1}]$ and the corresponding crack length intervals $[a_i, a_{i+1}]$ for the whole specimen life-time $(i = 1, 2..., n)$:

$$\int\limits_{a_{N_i}}^{a_{N_{i+1}}} \left[F\left(\frac{x}{b}\right)\sqrt{\pi x}\right]^{-m} dx = C(\Delta\sigma)^m \int\limits_{N_i}^{N_{i+1}} [1 + \xi(t)]dt. \tag{5.7}$$

Then we can write down the above equation in the following form:

$$\left[\Phi(a_{N_{i+1}} + a_{N_i})\right]^{-m} (a_{N_{i+1}} - a_{N_i}) = C(\Delta\sigma)^m (N_{i+1} - N_i)\eta_{i,i+1}, \tag{5.8}$$

where $\eta_{i,i+1}$ is a Gaussian random variable with

$$E\{\eta_{i,i+1}\} = 1, \qquad \mathrm{Var}\{\eta_{i,i+1}\} = \frac{\delta}{N_{i+1} - N_i}, \tag{5.9}$$

and

$$\Phi(a_{N_{i+1}} + a_{N_i}) = F\left(\frac{a_{N_{i+1}} + a_{N_i}}{2b}\right)\sqrt{\pi\frac{a_{N_{i+1}} + a_{N_i}}{2}}. \tag{5.10}$$

Calculating the natural logarithm (logarithm to base e) of the integrated crack growth equation (5.8), we obtain the following:

$$\ln(a_{N_{i+1}} - a_{N_i}) - \ln(N_{i+1} - N_i) = \ln\left[\Phi(a_{N_{i+1}} + a_{N_i})\Delta\sigma\right] m + \ln C + \zeta_{i,i+1}. \tag{5.11}$$

Now, using the experimental measurements $(a_{N_i}, N_i), i = 1, 2, ..., n$, we want to estimate the model parameters m and $\ln C$. Since the model is linear with respect to these parameters, we must adopt the method of linear estimation presented in Sec. 4 for equation (5.11). We identify the terms in equation (5.11) as:

$$x_i = \ln(a_{N_{i+1}} - a_{N_i}) - \ln(N_{i+1} - N_i), \tag{5.12}$$

$$A_{i1} = \ln\left[\Phi(a_{N_{i+1}} + a_{N_i})\Delta\sigma\right], \tag{5.13}$$

$$A_{i2} = 1, \tag{5.14}$$

$$\alpha_1 = m, \tag{5.15}$$

$$\alpha_2 = \ln C. \tag{5.16}$$

In the above we have assumed that random fluctuations of the crack length increments are small in comparison with the crack length, and the coefficients A_{ik} can be considered as deterministic constants. Moreover, for simplicity, we assume that the random variables representing the growth disturbance (noise)

$$e_i = \zeta_{i,i+1}, \tag{5.17}$$

are Gaussian with a zero mean and with equal variances σ^2. In the formulation of the model, in formula (5.9), we have assumed that the variances of the noises are of the form:

$$\mathrm{Var}\left\{\eta_{i,i+1}\right\} \approx \frac{\delta}{N_{i+1} - N_i}. \tag{5.18}$$

We know that, under realistic values of the numbers of cycles N_i, these variances are small and the denominators $N_{i+1}-N_i$, in (5.18) do not differ too much for all i. Therefore we can assume that the variances of random variables $\zeta_{i,i+1} = \ln \eta_{i,i+1}$ are for all i (approximately) equal:

$$\mathrm{Var}\left\{\zeta_{i,i+1}\right\} \approx \sigma^2 \tag{5.19}$$

and, moreover, the distribution of $\zeta_{i,i+1}$ can be approximately considered to be Gaussian.

6. Experimental data and estimation of the model parameters

As it is seen from the previous section, the parameters to be estimated in our simplified stochastic crack propagation model are m and $\ln C$. Now we must construct the numerical procedure of the parameter identification. We know that m and $\ln C$ are random variables and the algorithm must take this fact into account. Therefore we apply the statistical method of conditioning [7] for this model. This means that our procedure of identification of the statistical distribution of the pair $(m, \ln C)$ will be performed in the following two steps.

STEP 1. We consider the trajectory of the stochastic crack growth for the fixed elementary event $\omega' \in \Omega$. We assume, that this trajectory is governed by the Paris-Erdogan randomised equation (5.1) with the parameters $(m(\omega'), \ln C(\omega'))$. Using the crack growth model defined in Sec. 5 and the parameters estimation schedule from Sec. 4, we calculate the numerical values of the parameters $(m(\omega'), \ln C(\omega'))$.

STEP 2. We repeat the procedure of Step 1 for all the trajectories collected at the experiment (observed elementary events $\omega_i \in \Omega$)) obtaining the set of pairs $(m(\omega_i), \ln C(\omega_i))$, for $\omega_i \in \Omega$. Using the estimated values of the parameters $(m(\omega_i), \ln C(\omega_i))$, we identify the probabilistic distribution of the two-dimensional random variable $(m, \ln C)$.

REMARK. Let us remark that if the above procedure is applied for estimation of the value of the parameter $C(\omega)$ (or its mean value), then the proposed algorithm introduces some additional error of estimation. It is connected with this fact that

$$E(\ln C(\omega)|\text{measurements}) \neq \ln E(C(\omega)|\text{measurements}), \tag{5.20}$$

what means that the distributions (and, what it follows, the moments) of two random variables: the estimated value of $\ln C(\omega)$ and the random variable being the logarithm of the estimated value of $C(\omega)$ – are not equal. The difference of the above distributions is quite small if the variance of the estimated parameter $C(\omega)$ of the model is small. Finally let us remark that in our method of validation of the experimental data we use only one of the parameters ($\ln C(\omega)$, not $C(\omega)$), so we avoid a danger of inaccuracy caused by non-linear transformation of distributions.

7. Modelling stochastic crack growth using experimental data

The experiment of measurement of the stochastic crack growth is very complicated. It requires rigorous preparation of the material samples, exact repetition of excitations, environmental conditions, etc. Therefore in the literature one can find only a few papers where such data is presented. The examples of such results can be found in [4] and [13].

In our paper, as a material for the practical illustration of the above theoretical considerations, we use the Virkler experimental data of stochastic crack growth under periodic loading [13]. The results of this experiment are shown in Fig. 1. The authors performed the experiment for 68 samples of material, obtaining the trajectories of crack growth, each containing 164 measurement points. The experiment has been performed for the 2024-T3 aluminium alloy. The dimensions of all the samples were: length $a_{\text{tot}} = 558.8$ (mm), width $b = 152.4$ (mm) and thickness $d = 2.54$ (mm). The length of the fatigue crack was observed in the interval $9.00 \leq a \leq 49.8$ (mm); the stress intensity during the experiment was $\Delta\sigma = 48.28$, and the sinusoidal excitation frequency was 20 Hz.

The experimental trajectories are the fundamental basis for identification of the model parameters. To perform the procedure, we apply the algorithm proposed in Sec. 6, performed in two steps. In the first step we identify parameters $(m, \ln C)$ for each of the 68 trajectories of the stochastic crack growth. The estimated values of the parameter pairs are presented in Fig. 2.

It is seen that the parameters m_i and $\ln C_i$ are, with high accuracy, linearly dependent on each other. This means that in the second step of identification of the model, it is sufficient to consider only one parameter of the pair. Following the literature [3], we assume the normal distribution of the random variables $m(\omega)$ and $\ln C(\omega)$. This means that, in order to know the distributions, it is

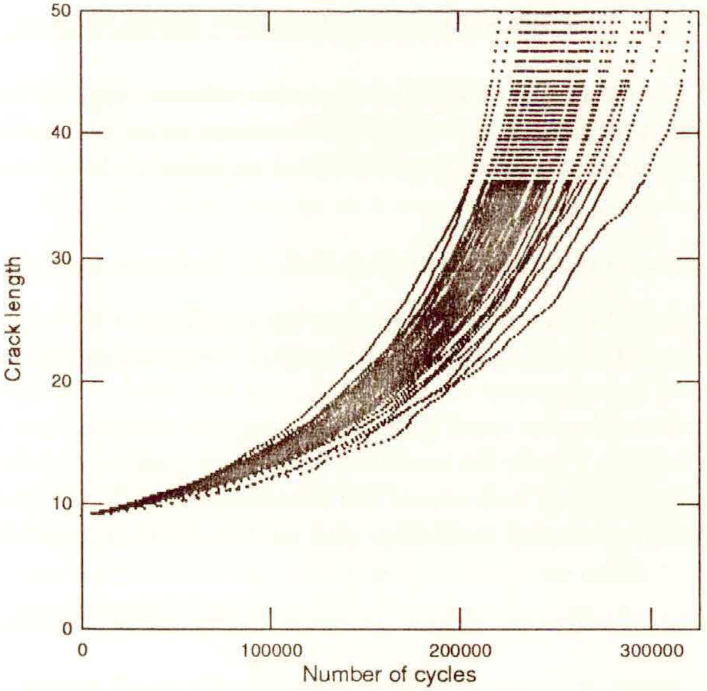

FIG. 1. Trajectories of the stochastic crack growth (results of the Virkler experiment).

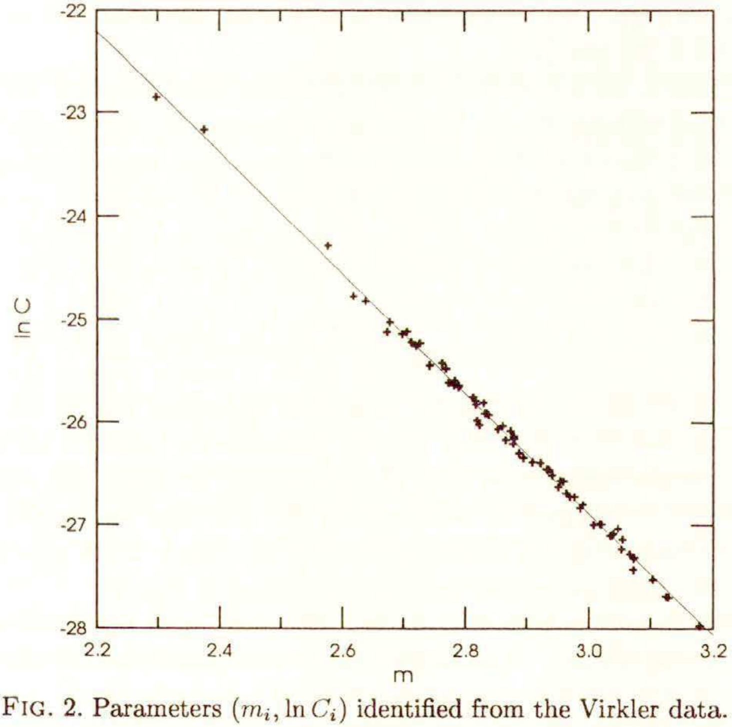

FIG. 2. Parameters $(m_i, \ln C_i)$ identified from the Virkler data.

enough to calculate their mean values and variances. In the second step of the conditioning procedure we estimate the moments of the parameter m according to the maximum likelihood estimators:

$$E\{m\} = \frac{1}{N}\sum_{i=1}^{N} m(\omega_i), \tag{7.1}$$

$$\mathrm{Var}\{m\} = \frac{1}{N}\sum_{i=1}^{N}(m(\omega_i) - E\{m\})^2. \tag{7.2}$$

Since we have observed the linear dependence of the parameters m and $\ln C$:

$$\ln C = Am + B, \tag{7.3}$$

to complete the identification of the model we should calculate the coefficients A, B, using the formula (4.13) for the linear estimator, and the experimental data presented in Fig. 2. The obtained moments of the random variables $m(\omega)$ and $\ln C(\omega)$ and the values of the parameters A and B are:

$$E\{m\} = 2.874, \tag{7.4}$$

$$\mathrm{Var}\{m\} = 0.02736, \tag{7.5}$$

$$A = -5.847, \tag{7.6}$$

$$B = -9.35, \tag{7.7}$$

$$E\{\ln C\} = AE\{m\} + B = -26.155, \tag{7.8}$$

$$\mathrm{Var}\{\ln C\} = A^2\mathrm{Var}\{m\} = 0.939. \tag{7.9}$$

8. Reliability of the experimental data and cross-validation

The procedure used for the identification of the model parameters needs the experimental data to obtain concrete numerical results. In our procedure we applied the data in two steps. In every step we performed the identification under an implicit assumption that the collected data are appropriate for our purpose. However, there is always a danger that this assumption cannot be justified. The general ideas concerning this fact have been presented in Sec. 1. Now we will show how the concrete example of estimation of the Paris-Erdogan model parameters on the basis of Virkler data, demonstrates the general idea of the cross-validation.

Let us discuss the results obtained in two steps of our conditioning procedure.

STEP 1. In this step we identify the sample parameters $(m_i, \ln C_i)$ for all 68 trajectories obtained in the experiment. For every trajectory we obtain a certain value of the parameters $(m, \ln C)$. To verify the validity of the estimated values, we try to reconstruct the Paris-Erdogan (deterministic or averaged) trajectories. The result of the calculation is presented in Fig. 3. During reconstruction of the trajectories we failed at 9 cases of 68 (9 times the sample paths with the identified parameters exploded before reaching the considered number of cycles). To explain this fact let us remark that (as it is visible in Fig. 1) some experimental trajectories of stochastic cracks are of the shape which is non-similar to the exponential Paris-Erdogan curve. Moreover, the length in time (number of cycles) is different for each experimental curve. Therefore the life-time of the modelled crack growing in the sample cannot be precisely determined. The discussion of analogous problems can be found in [8].

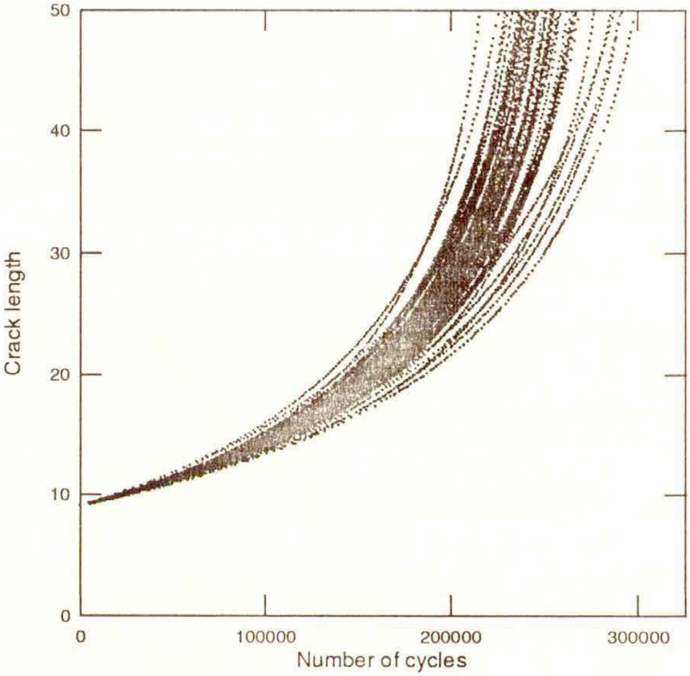

FIG. 3. Deterministic trajectories with parameters $(m_i, \ln C_i)$ estimated from the Virkler experimental data.

To study the effect of the trajectory length on the success of the procedure of the model parameters identification, we make the following calculations. We omit some number of the measurement points at the end of every curve in the procedure of Step 1. The results of such numerical experiment (the number of the identified pairs of the parameters for which the reconstruction of the Paris-Erdogan trajectory was impossible) are presented in the following table (the length of the trajectory is 164).

STEP 2. We estimate the model parameters (identify their distributions) basing on the data partially identified in Step 1. Now we try to verify the validity of the data for the complete identification procedure. We examine the reliability of the experimental data using the linear interdependence of two parameters in the Paris-Erdogan model of the stochastic crack growth. To do this, we compare the results of model identification obtained by two different methods.

Number of omitted data points on trajectory	Number of unsuccessful identifications
0	9
10	10
20	12
30	16
40	20
50	25
60	37
70	43
80	51

Assume that the value of the parameter m_i for fixed i is known (it is identified in the procedure of Step 1). Now we can calculate the values of the parameters A and B in the linear dependence (7.3).

METHOD 1. In this method the coefficients A and B are identified according to the formulae of Sec. 4 with the use of all the pairs of the estimated values $(m_i, \ln C_i)$.

METHOD 2. In this method the coefficients A and B are identified with the use of all the pairs of $(m_j, \ln C_j)$ except for the i-th pair.

Now, having the values of A and B estimated, we are able to calculate (according to (7.3)) the approximate value of the model parameter $\ln C_i$ for every m_i.

The first performed test shows, what is the influence of the i-th measured trajectory on the approximation quality of $\ln C_i$. Figure 4 shows the result of classical (one-point) cross-validation of the experimental data. The points on the plot marked with crosses represent the value of mean-square error of the approximation of the value of $\ln C_i$ estimated from the trajectory by $\ln C_i = Am_i + B$, where the parameters A and B were calculated by the Method 1. Points marked with circles represent the analogous error but for parameters A and B calculated according to the Method 2. It is seen that the differences in the approximation errors are significant for 9 measurements. This means that 9 measurements are not appropriate for the identification of the parameters of the Paris-Erdogan model. They contain a lot of information specific for themselves but useless for approximation of the general properties of the model.

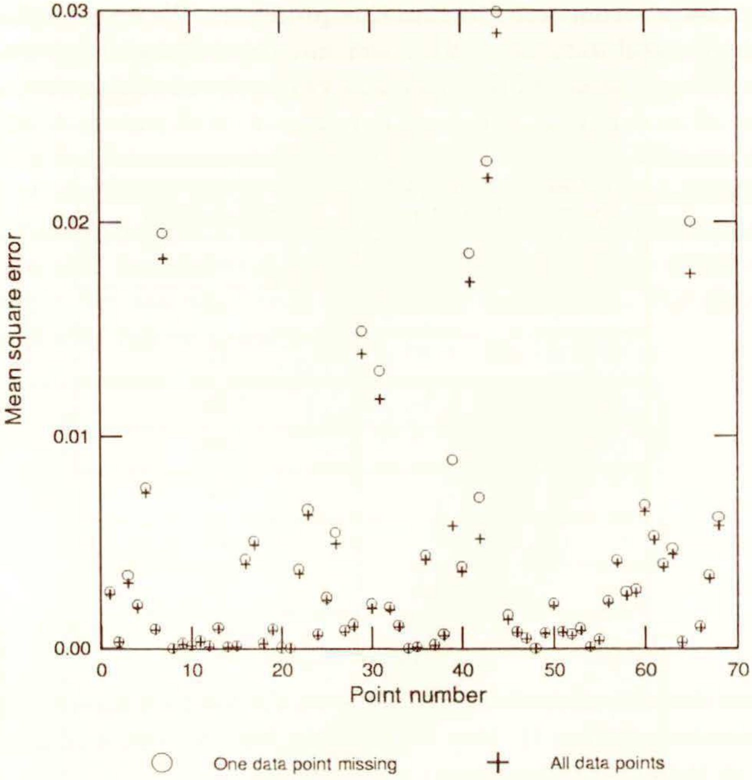

FIG. 4. The mean square error for approximation of the parameter $\ln C_i$.

The following identification method treats the cross-validation problem more generally.

METHOD 3. In this method, the coefficients A and B are identified with the use of all the pairs of the estimated values $(m_j, \ln C_j)$ except the k randomly selected pairs.

The results of the Method 3 are presented in Fig. 5. There are 3 lines in the plot. The dashed line shows the value of the mean square error of the approximation of the parameter $\ln C_i$, with the value m_i and formula (7.3), where the constants A and B were calculated according to the Method 2 (this is the sum of the errors for all 68 experimental trajectories). The solid lines show the analogous error but when the coefficients A and B are calculated according to the Method 3. The functions depend on k, the number of the omitted points (for two different random selections).

It is seen that, in general, omission in the approximation procedure of $\ln C_i$, at a given point just the measurement made at this point, gives the effect comparable to neglecting more than 30 randomly selected points (that is about 50% of the points considered in the estimation procedure). This means that each curve of the Virkler data is strongly informative for the estimation of the value of the parameters calculated for this curve.

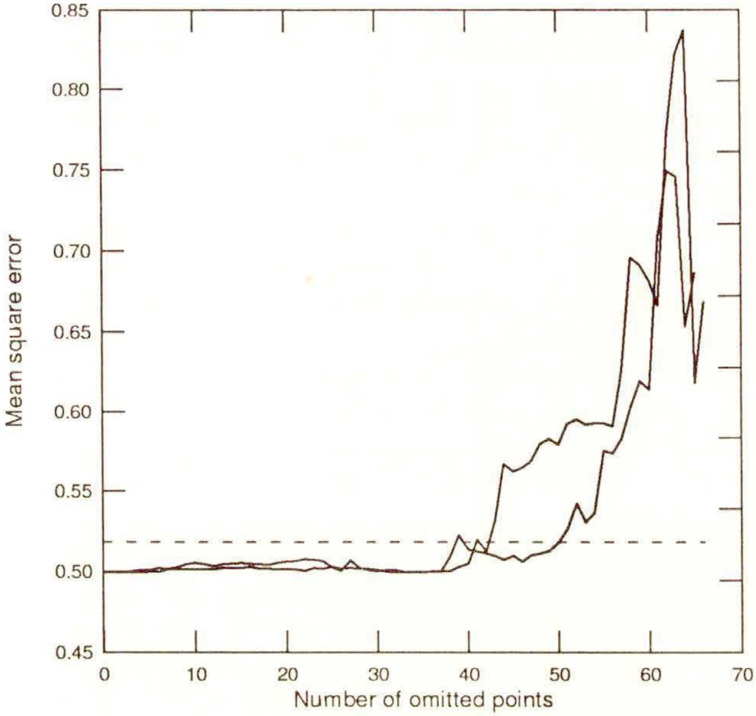

FIG. 5. The averaged mean square error of estimation of the parameter $\ln C_i$.

9. Closing remarks

One of the most important tasks of the experiment's design is the verification of the consistency of the measured experimental data. To analyse the data, we have applied the method analogous to the statistical procedure of cross-validation. Since the results of measurement had to be applied for identification of the parameters of a certain mathematical model, we applied this model (or, more precisely its parameters) as the quality measure of the set of experimental data. Such a methodology is very intuitive: the collected data can be more appropriate for one model, less appropriate or useless for another. The reasons for this fact can be very different. It can happen that some model is not adequate for description of the observed physical phenomenon and this fact must be always taken into account in the identification process. However, this is not the only reason of failure of the procedure. Sometimes the algorithms of the model parameters estimation require a specific structure of data. Therefore one must carefully design the experiment planning its duration, sampling in time, location of sensors over the sample, etc., taking into account the final destination of the obtained data. Summing up, validation of the experimental data must be always connected with the model where the data are utilised.

In this paper we have considered the following practical problem: for a given set of experimental data (Virkler data on the fatigue-crack length) and the mathematical model of a physical phenomenon (Paris-Erdogan randomised model of fatigue-crack growth), verify the validation of the data for identification of the model parameters. The conclusions regarding possibility of application of the Virkler data in the Paris-Erdogan model are the following:

- Virkler data applied in identification of the Paris-Erdogan randomised model are sensitive to the length in time (duration) of the sample trajectories. They are also very sensitive to omitting the results of certain sample measurements in the identification procedure.
- After the cross-validation procedure applied to the Paris-Erdogan equation, we must say that while the model gives a good qualitative description of the stochastic crack growth, there is a small possibility of prediction of the behaviour of the crack in a certain sample of a material. To estimate the parameters of certain trajectory with good accuracy, we should include into our calculations the experimental results obtained just for this trajectory.
- In the experiments of a kind analogous to the Virkler one, the number of the measured samples and the length of the observed trajectory is essential for the quality of identification of any mathematical model of the tested phenomenon.

To conclude our considerations we must say that while every experiment, before it is made, must be carefully designed, then the following cross-validation procedure can strongly confirm the applicability of the obtained data for mathematical modelling. This procedure indicates in particular the coherence of the obtained experimental data and the applied theoretical model of the phenomenon.

Acknowledgements

I wish to thank Professor KAZIMIERZ SOBCZYK for pointing the problem and a support during preparation of this paper.

References

1. P. CHAUDHURI, A. DEWANJI, *On a likelihood-based approach in nonparametric smoothing and cross-validation*, Statistics & Probability Letters, **22**, 7–15, 1995.
2. O. DITLEVSEN, *Random fatigue crack growth – a first passage problem*, DCAM, Rep.No. **298**, 1985.
3. O. DITLEVSEN, R. OLESEN, *Statistical analysis of the Virkler data on fatigue crack growth*, Engn. Fracture Mechanics, **25**, 2, 177–195, 1986.

4. H. Ghonem, S. Dore, *Experimental study of the constant-probability crack growth curves under constant amplitude loading*, Engn. Fracture Mechanics, **27**, 1, 1–25, 1987.

5. D. K. Hildebrand, J. D. Laing, H. Rosenthal, *Prediction analysis of cross classifications*, J.Wiley & Sons, New York 1977.

6. D. D. Joshi, *Linear estimation and design of experiments*, John Wiley and Sons, New York, New Delhi 1987.

7. Z. Kotulski, K. Sobczyk, *Effects of parameter uncertainty on the response of vibratory systems to random excitation*, J. Sound and Vibration, **119**, 1, 159–171, 1987.

8. C. J. Lu, W. Q. Meeker, *Using degradation measures to estimate a time-to-failure distribution*, Technometrics, **35**, 2, 161–174, 1993.

9. C. Radhakrishna Rao, *Statistics and truth*, Council of Scientific and Industrial Research, New Delhi 1989.

10. K. Sobczyk, B. F. Sencer, Jr., *Random fatigue: From data to theory*, Academic Press, 1992.

11. M. Stone, *Cross-validatory choice and assessment of statistical predictions*, Proc. Royal Statist. Soc. of London B, **36**, 2, 111–147, 1974.

12. M. Stone, *Cross-Validation: A Review*, Math.Operat.-Forschung Statist., Ser. Statistic, **9**, 127–139, 1978.

13. D. A. Virkler, B. M. Hillberry, P. K. Goel, *The statistical nature of fatigue crack propagation*, Transactions of the ASME – Journal of Engineering Materials and Technology, **101**, 148–153, 1979.

POLISH ACADEMY OF SCIENCES
INSTITUTE OF FUNDAMENTAL TECHNOLOGICAL RESEARCH
e-mail: zkotulsk@ippt.gov.pl

Received December 8, 1997.