Available online at www.eccomasproceedia.org
Eccomas Proceedia UNCECOMP (2017) 598-604

ECCOMAS
Proceedia

UNCECOMP 2017
2nd ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Rhodes Island, Greece, 15–17 June 2017

THE NON-PARAMETRIC APPROACH TO THE QUANTIFICATION OF THE UNCERTAINTY IN THE DESIGN OF EXPERIMENTS MODELLING

Jacek Pietraszek[1], Renata Dwornicka[1], Mariusz Krawczyk[2], and Maciej Kołomycki[1]

[1] Department of Software Engineering and Applied Statistics, Cracow University of Technology
37 Jana Pawła II, 31-864 Kraków, Poland
e-mail: pmpietra@mech.pk.edu.pl, dwornick@mech.pk.edu.pl, mkolomycki@gmail.com

[2] Chair of Information Systems and Computer Modeling, Cracow University of Technology
37 Jana Pawła II, 31-864 Kraków, Poland
e-mail: mariusz.krawczyk@mech.pk.edu.pl

Keywords: Non-Parametric Statistical Methods, Design of Experiments, DoE, Factorial Approach, RSM Approach, Uncertainty Quantification.

Abstract. *The classic design of experiments (DoE) typically uses the least-square method for a model identification and requires associated assumption about the normality of a noise factor. It is very convenience because it leads to a relative simple computations and well-known asymptotic statistics based on the normality assumption. However, if that assumption is not satisfied it may fail and obtained results may differ radically from the verification tests. The rationale for the caution may be the comparison of interval plots (based on the normality hypothesis) and box-plots (based on raw data). The useful approach is the bootstrap-based methodology which replaces the requirement of the normality assumption with weaker requirement of the independent and identical distribution (i.i.d.) of the random term. The industrial applications of this approach are still rare because the industry is very conservative and usually utilizes old well-known methods and typical numerical software like e.g. Statistica, Statgraphics or Minitab. This paper presents the bootstrap modeling of the random uncertainty in the two cases: the factorial designed experiment and the response surface experiment.*

1 INTRODUCTION

Methods of the approximation and the prediction rapidly evolved in recent years. New approaches came from new branches of the statistics and the artificial intelligence area: non-parametric, data-driven, stochastic etc. Apart from this, the classic approach of the design of experiment methodology (DoE) is still very useful and popular in the industry full-scale production as well as in the R&D laboratories.

Peer-review under responsibility of the organizing committee of UNCECOMP 2017.
doi: 10.7712/120217.5395.17225

The classic design of experiments (DoE) utilizes factorial model of fixed effects since R.A. Fisher introduced the analysis of variance (ANOVA) [1] and the latin squares [2] in mid-twenties of the 20th century, F. Yates proposed two-level fractional factorials [3] in mid-thirties and G. Taguchi orthogonal arrays [4] in sixties. Simultaneously, the DoE utilizes the response surface methodology (RSM), however usually only in the simple form of low-degree polynomials, since G.E.P. Box introduced RSM [5] in early fifties of the 20th century and H. Scheffé [6] extended this methodology into the mixture designs in late fifties.

The typical procedure of the model identification assumes the normal distribution of the data noise [7] and uses the least square method [8] to identify parameters of the model with *a priori* assumed structure what set this approach in the group of parametric methods. The terms of a model are repeatedly eliminated in the specific backward stepwise regression, while three indicators: (a) the least significance of parameters, (b) the significance of the lack of fit and (c) the conformity of residuals with the normal distribution are simultaneously observed to make a decision to stop or to continue the elimination procedure.

Practically, for small datasets, the conformity with the normal distribution has very weak reliability and it leads to the very uncertain assessment of parameters statistical significance and bounds of their confidence intervals. The bootstrap approach [9] with simulation-based identification of parameters confidence intervals appears to be better solution than theoretically proved but only asymptotically equal *t*-distribution [7]. It seems to be specifically more important for the non-parametric models where the model structure is created dynamically during data analysis e.g. artificial neural networks , regression trees etc., while theoretically-based distribution cannot be determined, because their final structures are so non-linear that analytical identification of associated distributions and their confidence intervals is not possible.

The key issue in the bootstrap approach is to make proper identification of the random term inside the model structure because the i.i.d. requirement have to be satisfied by this term.

2 METHODS

DoE methodology splits into two branches: the factorial approach [10, 11] for qualitative input variables (factors) and the response surface methodology (RSM) [11, 12] for quantitative input variables. Obviously, the models for mixed inputs, qualitative and quantitative, may be identified however it is not so easy.

The specific models used in these methodologies and the location of the random term are described in subsequent chapters in details.

2.1 Factorial model

The factorial approach is typical for qualitative input variables. Their settings formally are labels, not numerical values, in contrast to quantitative ones. The model predicts response for assumed settings and its typical structure is the fixed-effects model [7] with additive terms which depends on single factors (main effects, linear effects), two factors (second-order interaction), three factors (third-order interaction) and eventually up to the highest-order interaction which merges all factors contained in the model. So rich model requires symmetrically equivalent rich experimental design: all combinations of all factor's levels have to be tested experimentally.

Such design is known as *full factorial*. It may be described by implication: if you need all possible interactions, you have to test all possible combinations of levels. It is clear that reverse question has to be ask: could I omit tests of some combinations of levels, if I needn't all

interactions? The answer is positive and was given first by Fisher [2] in the form of *latin squares* and later by Yates [3] who developed two-level *fractional factorials*.

The sample structure of the fixed-effects model with three factors and terms up to the third-order interaction is defined as following:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon \tag{1}$$

where:

y_{ijk} – the random variable describing the predicted output,

μ – the average output,

α_i – the linear (main) effect of the first factor α at its level i,

β_j – the linear (main) effect of the second factor β at its level j,

γ_k – the linear (main) effect of the third factor γ at its level k,

$(\alpha\beta)_{ij}$ – the second-order interaction effect of the factors α and β at their levels i and j,

$(\alpha\gamma)_{ik}$ – the second-order interaction effect of the factors α and γ at their levels i and k,

$(\beta\gamma)_{jk}$ – the second-order interaction effect of the factors β and γ at their levels j and k,

$(\alpha\beta\gamma)_{ijk}$ – the third-order interaction effect of the factors α, β and γ at their levels i, j and k,

ϵ – the random term describing impact of all uncontrolled noise factors; typically the normal distribution with an unknown variance $N(0, \sigma^2)$ is assumed, however it is only a hypothesis.

The equation (1) shows the general form of the stochastic model, but the numerical simulation like e.g. the bootstrap requires replacement of the random variable with its particular observations. It leads to smoothly different relationship:

$$y_{ijk|r} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + r_{ijk|r} \tag{2}$$

where:

$y_{ijk|r}$ – the output measured at levels i, j, k of the respective factors and at r repetition of the test,

$r_{ijk|r}$ – the residuum (difference between the real measurement of output and its prediction) observed at levels i, j, k of the respective factors and at r repetition of test; it should be noted that r symbol has two difference means: the residuum and the index of test repetitions (in the subscript).

Just the set of observed residuums $\{ r_{ijk|r} \}$ will be the source for the future subsampling in the bootstrap approach. The algorithm's steps are as following [9]:

a) the dataset of measurements $\{ y_{ijk|r} \}$ is used to the identification of effects (see Eq.1) typically by the least-squares method,

b) the identified effects are used to predict output $\{ \overline{y}_{ijk} \}$ for those levels i, j, k at which measurements were obtained,

c) the dataset of residuums $\{ r_{ijk|r} \}$ is created based on differences between the observed measurements and the predicted outputs, according to the formula:

$$r_{ijk|r} = y_{ijk|r} - \overline{y}_{ijk} \tag{3}$$

d) the bootstrap subsampling is made by a random draw from the dataset of residuums $\{ r_{ijklr} \}$, added subsequently to the predicted outputs, resulting in the dataset of bootstrapped outputs:

$$\hat{y}_{ijklr} = \overline{y}_{ijk} + draw(r_{ijklr}) \tag{4}$$

e) the bootstrapped outputs are used to identification of new estimates of model effects.

The steps (d) and (e) should be repeat many times (at least thousands times) resulting in a dataset of enough size to evaluate necessary statistics.

2.2 Response surface model

The response surface approach is typical for quantitative input variables. Their settings are continuous numerical values. The model predicts response for assumed settings and its typical used structure is a second-order polynomial [7] being a local example of Taylor series. Such model is a specific modification of a classic approximation problem due to a random term describing an impact of all uncontrolled noise factors:

$$y = f(x_1,\ldots,x_i;b_0,b_1,\ldots,b_j) + \epsilon \tag{5}$$

where:

y – the random variable describing the predicted output,

f – the assumed function, usually first or second-order polynomial,

x_i – i-th input variable,

b_j – j-th model parameter,

ϵ – the random term describing impact of all uncontrolled noise factors; typically the normal distribution with an unknown variance $N(0, \sigma^2)$ is assumed, however it is only a hypothesis.

Assumption of f function as the first or the second-order polynomial leads to very convenience situation: all model parameters are linearly set and may be identified by the least-squares method.

Further procedure has to be split. If a dataset of the output is collected in a designed experiment with strictly controlled input variables then the random term is ε from stochastic Eq.5 is replaced with particular observed residuum r and the dataset of residuum will be the source for bootstrap draw. The next steps of the bootstrap approach are exactly the same a previously mentioned for the fixed-effect model.

But if the dataset of output is collected in a passive experiment and input variables are loosely controlled (if any), then a point from multidimensional space of observations $(x_1,\ldots, x_i; y)$ is established as a the random element. The algorithm's steps are as following [9]:

a) the dataset of input values (formally input, but only observed in passive mode) and observed measurements $\{ (x_1,\ldots,x_i; y)_j \}$ is used to the identification of the basic model parameters (see Eq.5) typically by the least-squares method,

b) the bootstrap subsampling is made by a random draw from the dataset of input values and observed measurement $\{ (x_1,\ldots,x_i; y)_j \}$ resulting in a bootstrapped dataset of the same size $\{ (\hat{x}_1,\ldots,\hat{x}_i; \hat{y})_j \}$, and a bootstrapped model with parameters $\hat{b}_k$ is identified:

$$\sum \left[\hat{y}_{1j} - f(\hat{x}_{1l\,j},\ldots,\hat{x}_{il\,j};\hat{b}_k) \right]^2 \longrightarrow \min \tag{6}$$

c) the bootstrapped model and its output are used to identification of new estimates of model effects.

The steps (b) and (c) should be repeat many times (at least thousands times) resulting in a dataset of enough size to evaluate necessary statistics.

3 CASE STUDIES

Two case studies are presented for a bootstrap-based non-parametric estimation:

– estimation of the parameter's significance for a fixed-effects model,
– estimation of the parameter's significance for a response surface model in passive mode experiment.

3.1 Case study for a fixed-effects model

The investigation was conducted on the materials science matter: the analysis of of the ceramic shell mould of the airfoil blade casting [13]. One casting was selected for the microstructural and statistical analysis, and finally cut off. The cross-sections were included and prepared as metallographic samples from nickel-based superalloy. To reveal the microstructure of the investigated material the surfaces of the samples were etched, observed by a scanning electron microscope and obtained images analyzed by a computer-aided image analysis program to estimate quantitatively the main parameters describing the (γ+γ') eutectic islands that occurred in the investigated superalloy. Details of the bootstrap-based simulation provided on the mentioned data are available in [13, 14] and here a briefly summary of analysis and results is included.

Dataset was created as six groups containing areas of eutectic islands identified and measured for each of six traces. The primary selected tool for the test of the traces homogeneity was classic ANOVA, but unfortunately the diagnostic test rejected the variances homogeneity, what is the basic assumption in ANOVA.

As an alternative method, the bootstrap-based identification of confidence intervals was provided. The one-way general linear model with six levels was identified to construct the residuum dataset, which is necessary to make random draw in the bootstrap. Next, the bootstrapped models were identified in 10 thousand iterations and parameters were collected to construct associated statistics (Figure 1).

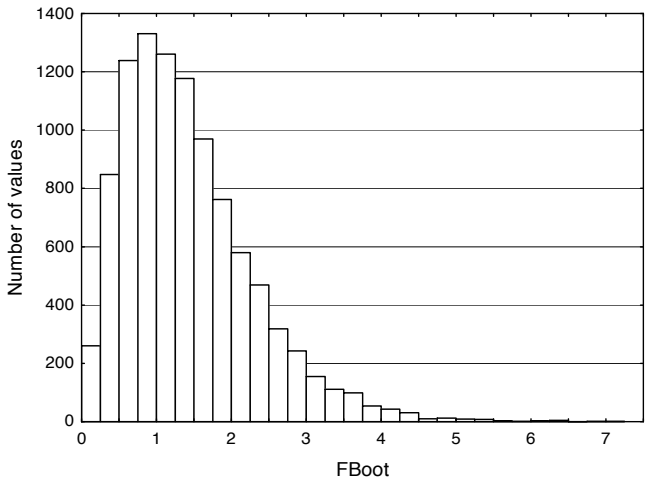

Figure 1: Distribution of the bootstrapped F statistics (source: [13])

The bootstrap simulation experiment revealed that ANOVA results are rather uncertain. The bootstrapped is approximately equal to 0.75 while classic ANOVA *p*-Value obtained from experimental data gave 0.84. It leads to the conclusion that criterion *p*-Value should be treated more as a fuzzy basis for the decision, not the sharp one just on 0.05.

3.2 Case study for a response surface model in a passive mode experiment

The case study [15] used data obtained during investigation on compression vertebral fractures prediction based on computer tomography (CT) and microtomography (μCT) images. The significant difference in a resolution between these two class of images leaded to different prediction models. The small sample size (23 compressed and scanned vertebraes) and the high dimensionality of detected properties imposed the necessity of an alternative approach to the analysis, other than classic one derived with a requirement of the normality.

The RSM model was used to construct the prediction model to fit the observed strength of vertebraes with the crushing force. The bootstrap method was used to evaluate confidence intervals for effects. Finally, the existence of zero inside the intervals was inspected. If any of intervals contained zero, the null hypothesis of was rejected i.e. the parameter was treated as statistically insignificant.

The source dataset contained 23 records of three variables selected as predictors (input variables) and measured strength selected as output. The number of draw iterations was set to 10.000 to easy selection of quantiles from the bootstrapped dataset. After the full bootstrap procedure, the descriptive statistics were evaluated for model parameters. The bounds of the confidence intervals were easy identified due to the selected number of bootstrap iterations. They were values found at positions 250 and 9750 in the sorted bootstrapped results. Similarly, the bootstrapped *p*-Value was evaluated as relative position of sign switching inside the sorted bootstrapped results.

The obtained results led to the same decision: all linear coefficients significant and a intercept insignificant, but bootstrap-based results showed stronger significance of the coefficient than classic results.

4 CONCLUSIONS

- The non-parametric analysis based on the bootstrap approach allows to conduct effective analysis in such situations, when requirements of classic assumptions are not met.
- The bootstrap-based analysis requires proper identification of the i.i.d. random term (residuum vs. multidimensional point) but its implementation is easy and does not require additional subtle assumptions.
- The non-parametric bootstrap-based analysis is useful for processing data with a weak assessment of the distribution e.g. small datasets.
- The non-parametric bootstrap-based analysis may be easy automatized for processing huge or even big data, which distributions do not fit into any theoretical distribution.
- Further research will be conducted, especially in full-scale industry environment [16, 17, 18], because of engineers expectations about automatized and easy-to-interpret analysis methods.

REFERENCES

[1] R.A. Fisher, *Statistical Methods for Research Workers*. Oliver and Boyd Press, Edinburgh, 1925.

[2] R.A. Fisher, The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture of Great Britain* **33**, 503-513, 1926.

[3] F. Yates, Complex experiments. *J. Roy. Statist. Soc.* **Suppl.2**, 181-247, 1935.

[4] M.S. Phadke, *Quality Engineering Using Robust Design*. Prentice Hall International, Inc., London, 1989.

[5] G.E.P. Box, K.B. Wilson, On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **13**, 1-45, 1951.

[6] H. Scheffé, Experiments with Mixtures. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **20**, 344-360, 1958.

[7] P.W.M. John, *Statistical design and analysis of experiments*. SIAM, Philadelphia, 1998.

[8] A.J. Izenman, *Modern Multivariate Statistical Techniques. Regression, Classification and Manifold Learning*. Springer Science+Business Media, LLC, New York, 2008.

[9] J. Shao, D. Tu, *The Jackknife and Bootstrap*. Springer, New York, 1995.

[10] D.C. Montgomery, *Design and Analysis of Experiments*. John Wiley & Sons, Inc., Hoboken, 2008.

[11] O. Kempthorne, K. Hinkelmann, *Design and analysis of experiments. Vol.1. Introduction to experimental design*. John Wiley & Sons, Hoboken, NJ, USA, 2007.

[12] R.H. Myers, D.C. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. John Wiley & Sons, 1995.

[13] A. Szczotok, J. Nawrocki, A. Gądek-Moszczak, M. Kołomycki, The bootstrap analysis of one-way ANOVA stability in the case of the ceramic shell mould of airfoil blade casting. *Solid State Phenomena*, **235**, 24-30, 2015.

[14] J. Pietraszek, R. Dwornicka, A. Szczotok, The bootstrap approach to the statistical significance of parameters in the fixed effects model. M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.). *ECCOMAS Congress 2016 – Proceedings of the 7th European Congress on Computational Methods in Applied Sciences and Engineering*, Crete, Greece, June 5-10, 2016, 6061-6068.

[15] J. Pietraszek, L. Wojnar, *The bootstrap approach to the statistical significance of parameters in RSM model*. M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.). *ECCOMAS Congress 2016 – Proceedings of the 7th European Congress on Computational Methods in Applied Sciences and Engineering*, Crete, Greece, June 5-10, 2016, 2003-2009.

[16] J. Pietraszek, A. Gadek-Moszczak, T. Toruński, Modeling of errors counting system for PCB soldered in the wave soldering technology. *Advanced Materials Research*, **874**, 139-143, 2014.

[17] J. Pietraszek, E. Skrzypczak-Pietraszek, The optimization of the technological process with the fuzzy regression. *Advanced Materials Research*, **874**, 151-155, 2014.

[18] J. Pietraszek, E. Skrzypczak-Pietraszek, The uncertainty and robustness of the principal component analysis as a tool for the dimensionality reduction. *Solid State Phenomena*, **235**, 1-8, 2015.