TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
MECHANICS | MECHANIKA

1-M/2013

JACEK PIETRASZEK*, ANETA GĄDEK-MOSZCZAK*

THE CONCEPT OF THE VARIANCE ESTIMATION FOR THE NEURAL NETWORK APPROXIMATOR BY JACKKNIFE SUBSAMPLING

KONCEPCJA ESTYMACJI WARIANCJI APROKSYMATORA NEURONOWEGO ZA POMOCĄ PODPRÓBKOWANIA JACKKNIFE

Abstract

The estimation of a variance for a semi-parametric neural network model variance for geometric properties of sintered metal will be done on the basis of jackknife subsampling method. Calculation results are of great practical significance because it will be possible to use proposed approach in similar microscale modelling. The proposed approach is simple and has many advantages if model identification procedure is computational expensive.

Keywords: *jackknife variance estimation*, *statistical methods*, *image analysis*, *error estimate*, *metal powder sintering*

Streszczenie

W artykule przedstawiono estymację wariancji półparametrycznego modelu neuronowego cech geometrycznych spieku metali przeprowadzoną za pomocą metody podpróbkowania *jackknife*. Obliczone wyniki są cenne z uwagi na możliwość zastosowania proponowanego podejścia do analogicznych zagadnień modelowania w mikroskali.

Słowa kluczowe: *estymator wariancji jackknife*, *metody statystyczne*, *analiza obrazu*, *estymacja błędu*, *spiekanie proszków metali*

* PhD. Jacek Pietraszek, PhD. Aneta Gądek-Moszczak, Institute of Applied Informatics, Faculty of Mechanical Engineering, Cracow University of Technology.

1. Introduction

1.1. Geometrical properties of 2D images of sintered metals

The ferritic-austenitic stainless steel was obtained by sintering the mixture of ferritic stainless steel AISI 434L powders with different amount of additions: Mn, Ni and Si. Effects of additions on quality of sintered products were studied [1]. Microscopic geometrical properties of porosity formulate one of the quality assessment criterion. In porous materials the character of the pore structure strongly effects on its mechanical properties. The microscopic structure of sintered samples was investigated by computer image analysis methods [2]. The obtained 2D optical image was processed by specialized software for image analysis [3]. The quantitative properties of pores were identified, among others pores area and circularity ratio. These properties were non-homogeneous and thus an empirical cumulative distribution and a histogram were their appropriate descriptions. For the evaluation of sinter produced with a variety of additions, there is necessary to create models for simulating the distributions of geometric properties including estimation of some measures of uncertainty e.g. variance.

1.2. Modelling

The empirical cumulative distribution function is stepwise and the histogram has little smoothness. For sintering simulation purposes, it is appropriate to create a smooth model of empirical cumulative distribution function with confidence bands [4, 5]. Unavoidable setting uncertainties of compacting and sintering lead to dispersion of obtained sinters characteristics. Reducing dispersion would be possible using replications i.e. compacting and sintering in the same setting nominal conditions. Replicated compacting and sintering, however, introduce into the experiment additional block factors associated with individual systematic differences in mixing additives, compacting and sintering, even at the same nominal settings. Identification of block factors impact would require a significant increase in the number of completed samples, metallographic specimens and analyses [6]. Instead, it is possible to use methods based on controlled perturbations of obtained model and analyse the impact of disturbances on the modelled output. The most promising non-parametric models do not allow to examine these effects with analytical methods, as is used in the classical perturbation theory [7, 8].

1.3. Jackknife method

Nonparametric and semi-parametric models have not imposed a priori regression formula [9]. The formula structure is adaptively data driven what allows much better fit prediction to the raw data. It should be noted that, contrary to commonly named, non-parametric models have parameters. In most cases, the identification of non-parametric models (neural networks, NPMEL, FEM) is computationally very expensive, so it is advisable to seek the most cost-effective use of the procedures for the identification of such models. In the absence of known in advance function formula, it is not possible to determine probability distributions of output variables and their confidence bands by analytical methods.

Numerical determination of such distributions is possible by using Monte Carlo methods [10–12]. The use of most frequently encountered bootstrap method [13], allowing to obtain the whole probability distribution of the output, would require multiple (from a few hundred to several thousand times) identifications of the auxiliary parametric models, which can be very expensive computationally. If the uncertainty assessment would be limited to estimates of variance, it will allows to implement jackknife subsampling procedure [14], which requires the use of a relatively small and acceptable (from a few to tens of times) number of non-parametric identification of auxiliary models. Simultaneously, the knowledge of the variance enables fully reliable estimate searched uncertainty.

2. Materials and methods

2.1. Sintered sample [1]

The water atomized powder AISI 434L of Höganäs Corporation [15] was used as a base powder. The additions were manganese powder, silicon powder and nickel powder. The four blends were prepared, but in this article, data collected for the sample 434 L+14% Mn are analysed.

2.2. Image analysis

The images of the sample 2D structure was acquired by Olympus camera model DP-25 coupled with Nikon microscope model Eclipse E400 and PC computer system. The images were analysed by algorithm written in environment of ADCIS Aphelion software [3] and pores were detected and quantified.

2.3. Neural network model

The basic model is an artificial neural network with feed-forward multi-layer architecture [16]. This network has single input and single output:

$$y = F(x, \beta_i, f_j) \tag{1}$$

where:

x – pore size,
y – predicted quantile corresponding to the size of a pore,
β_i – i-th weight of neural network synapse,
f_j – activation function of j-th neuron.

This basic model was identified (learned) by Statsoft Statistica program with Automatic Neural Network module. The best topology and activation functions were selected automatically by the program. Sub-sampling models were identified (learned) by Statsoft Statistica program with Automatic Neural Network module, but only synapses weight were selected while the topology and activation functions were the same as in the basic model.

2.4. Jackknife variance estimator for nonlinear regression

Statistic is a function of data selected according to some principle e.g. likelihood, sufficiency etc. Such statistic, prior to data collection, is a random quantity having probability distribution called the *sampling distribution*. The sampling distribution of a statistic depend on the underlying population and therefore is unknown. The *jackknife* is a method for estimating the sampling distribution of a statistic and its characteristic. The comprehensive elaborate is available at Shao and Tu [14]. The general nonlinear model of feed forward neural network has the following formula (eq. 2):

$$y_i = f(x_i, \beta) + \varepsilon_i, \qquad i = 1, \ldots, n \tag{2}$$

where:

- β – q-vector of unknown parameters (weights and biases in neural networks),
- f – known function nonlinear in β,
- x_i – p-vectors,
- ε_i – random errors with $E(\varepsilon_i|x_i) = 0$.

Pairs (y_i, x_i) are i.i.d. with a finite second moment. Typical estimators of weights and biases β in neural network regression problems are obtained in supervised approach. The estimators $\overset{\wedge}{\beta}_{LS}$ are taken as least squares estimator (eq. 3)

$$L_n(\overset{\wedge}{\beta}_{LS}) = \min L_n(\gamma), \quad \gamma \in B \tag{3}$$

where B is a set of all possible values of β and criterion L_n is of formula (eq. 4):

$$L_n(\gamma) = \frac{1}{2}\sum_{i=1}^{n}[y_i - f(x_i, \gamma)]^2 \tag{4}$$

If the parameter of interest ϑ is defined as a given function g of parameters β (eq. 5):

$$\vartheta = g(\beta) \tag{5}$$

then the LSE estimator of ϑ is (eq. 6):

$$\overset{\wedge}{\vartheta}_{LS} = g(\overset{\wedge}{\beta}_{LS}) \tag{6}$$

and jackknife variance estimator of $\overset{\wedge}{\vartheta}_{LS}$ is (eq. 7):

$$\upsilon_{\mathrm{JACK}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\overset{\wedge}{\vartheta}_{LS,i} - \frac{1}{n}\sum_{j=1}^{n}\overset{\wedge}{\vartheta}_{LS,j}\right)^2 \tag{7}$$

The values $\overset{\wedge}{\vartheta}_{LS,i}$ are based on $\overset{\wedge}{\beta}_{LS,j}$, being the LSE of β obtained after deleting the i-th pair (y_i, x_i). Now, if g is defined as f at arbitrary x i.e. (eq. 8):

$$g(\beta) = f(x, \beta) \tag{8}$$

then ϑ is mean value of f at x. With this assumption, $\overset{\wedge}{\vartheta}_{LS}$ and υ_{JACK} are jackknife estimators of the mean and the variance of the function f at arbitrary x (eq. 9):

$$\begin{aligned} E(\overset{\wedge}{\vartheta}_{LS}) &\xrightarrow[n\to\infty]{} E(y|x) \\ E(\upsilon_{\mathrm{JACK}}) &\xrightarrow[n\to\infty]{} \mathrm{var}(y|x) \end{aligned} \tag{9}$$

Now, evaluating estimators $\hat{\vartheta}_{LS}$ and υ_{JACK} at any arbitrary x, the mean and variance can be estimated. These estimators are used in the rest of the article.

2.5. General idea of simulation

The general outline of the workflow consists from the following 8 stages: (1) pre--processing of raw data from image analysis; (2) identification of the basic neural network model for the whole sample; (3) evaluating predictions from the basic neural network model; (4) decomposing of data into sub-samples; (5) identification of auxiliary neural networks models for sub-samples; (6) evaluating predictions from the auxiliary neural network models; (7) *jackknife* processing of the predictions from sub-samples models; (8) analysis of results.

2.6. Computational software

Image analysis was performed using ADCIS Aphelion package [3]. Numerical simulations were performed using PTC Mathcad version 15 [17]. Statistical analysis and significance tests were performed using Statsoft Statistica package [18].

3. Results

3.1. Raw data

The raw data obtained from the image analysis contain 17800 records. There were detected many pixelization artefacts associated with small objects below 32 pixel size. They were trimmed out with threshold of 32 pixels and data records were reduced to the number of 4544. Data were very irregular with lacks in many area sizes. The direct processing on this records is futile due to these irregularities generating only noise and smooth approximation is desirable. Next, area of pores were classified into 9 classes presented in Tab. 1. Class boundaries are arranged densely in these places, where the curvature is greater.

Table 1

Distribution of pores area frequency

Pore area [pixels]	32	43	62	93	123	188	265	634	2931
Number of pores with area less or equal	0	1000	2000	3000	3500	4000	4250	4500	4544

3.2. Neural models

The best fitted neural network model for pores area was topology 1-2-1 with *logistic* function for hidden layer and logistic function for output neuron. This topology was fixed as well as activation functions and subsample's identifications were processed. Predictions of full sample (basic) model, jackknife variance estimator and associated standard deviations are presented in Tab. 2.

Table 2

Neural network predictions and jackknife variance estimators for pores area frequency

Area	Measure	Full sample prediction	Residual	v_{jack}	$stdev_{jack}$
32	0	1142	1142	1585782	1260
43	1000	1315	315	1141451	1065
62	2000	1817	–183	385877	616
93	3000	2922	–78	846047	919
123	3500	3657	157	365754	602
188	4000	4207	207	23522	153
265	4250	4347	97	29215	171
634	4500	4405	–95	51505	227
2931	4544	4406	–138	46381	216

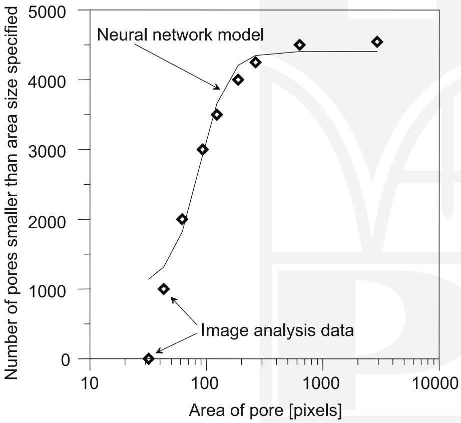

Fig. 1. The comparison of image analysis data and neural network model for distribution of pores area

Rys. 1. Porównanie rozkładów powierzchni porów dla danych z analizy obrazu oraz modelu neuronowego

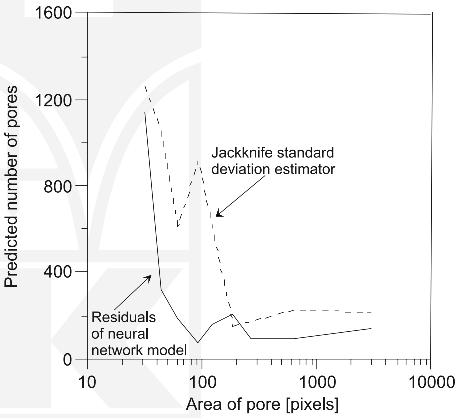

Fig. 2. The comparison of residuals and jackknife standard deviation estimator

Rys. 2. Porównanie wartości resztowych modelu neuronowego oraz odchylenia standardowego *jackknife*

Residuals may be treated as normally distributed (Kolmogorov test gives *p-value* = = 0.142). The comparison of image analysis data and neural network is presented on Fig. 1. The log scale for horizontal axis was selected due to saturation effects which make linear axis unreadable. The comparison of residuals and jackknife standard deviation estimator is presented on Fig. 2.

4. Discussion of the results

Neural network model approximates distribution of pores area continuously as expected. Model shows large errors at the left bound (for small pores). It does not reach the exact zero value but predicts the value of 1142 (see Fig. 1). This behaviour is reflected in both: residual values and an estimator of standard deviation predicted using *jackknife* method (Fig. 2). Estimator generally behaves in a similar way as real residual values of the model, but in one place (at pore size of 100) shows a sharp local spike, which looks like a slightly phase-shifted reflection of the smaller jump in the residuals. Such behaviour requires more study to determine whether it is associated with a specific model and data set, or is it the behaviour of the *jackknife* procedure interfering with a neural network model.

5. Conclusions

Metallographic studies were performed with sintered powder of stainless steel AISI 434L. Cumulative distribution of frequencies of the pore area were modelled by a neural network approximator and such approach produced a smooth waveform. The subsampling *jackknife* approach was used to estimate a variance and standard deviation of predicted values. Satisfactory results were achieved. This argues for the wider use of this methods with nonparametric approximator in the context of materials science.

References

[1] Sekuła M., *Sintering process analysis for modified powders of stainless steel AISI 434L*, Ph.D. Thesis., Cracow University of Technology, Kraków 2006.

[2] Wojnar L., *Image Analysis: Applications in Materials Engineering*, CRC Press, Boca Raton, 1998.

[3] Aphelion v 4.1.1., ADCIS, Saint-Contest, France 2012.

[4] *Springer Handbook of Engineering Statistics*, Springer, London 2006.

[5] Heinz S., *Mathematical Modeling*, Springer, Heidelberg 2011.

[6] Hinkelman K., Kempthorne O., *Design and Analysis of Experiments. Volume 2: Advanced Experimental Design*, John Wiley & Sons, Hoboken, NJ, USA 2005.

[7] Muralidhar K., Sarathy R., *A theoretical basis for perturbation methods*, Stat Comput 13, 2003, 329.

[8] Muralidhar K., Sarathy R., *An enhanced data perturbation approach for small data sets*, Decision Sci 36, 2005, 513.

[9] Horowitz J.L., *Semiparametric Models*, in *Handbook of Computational Statistics*, Gentle J.E., Härdle W.K., Mori Y. (eds.), Springer, Berlin Heidelberg, 2012, 597-618.

[10] Liu J.S., *Monte Carlo Strategies in Scientific Computing*, Springer Science+Business Media, LLC, New York 2008.

[11] Robert C.P., Casella G., *Monte Carlo Statistical Methods*, Second Edition, Springer Science+Business Media, LLC, New York 2004.

[12] Rubinstein R.Y., Kroese D.P., *Simulation and the Monte Carlo Method*, John Wiley & Sons, Hoboken, NJ, USA 2008.

[13] Mammern E., Nandi S., *Bootstrap and resampling, in Handbook of Computational Statistics*, Gentle J.E., Härdle W.K., Mori Y. (eds.), Springer, Heidelberg 2012

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

[15] *Höganäs Handbook for sintered components*, Höganäs AB.

[16] Samarasinghe S., *Neural Network for Applied Sciences and Engineering*, Taylor & Francis Group, LLC, Boca Raton, FL, USA 2007.

[17] Mathcad version 15, Parametric Technology Corporation, 140 Kendrick Street, Needham, MA 02494, USA, 2010.

[18] STATISTICA (data analysis software system), version 10., StatSoft, Inc., Tulsa, OK, USA 2011.