ADVANTAGES AND DISADVANTAGES OF VARIOUS UNCERTAINTY ASSESSMENT METHODS IN MATERIAL AND TECHNOLOGICAL PREDICTIVE MODELS

JACEK PIETRASZEK[1], MARIUSZ KRAWCZYK[2] AND ANNA KIEŁBUS[3]

[1]Cracow University of Technology, Faculty of Mechanical Engineering
Poland, 31-864 Kraków, Al. Jana Pawła II 37
jacek.pietraszek@mech.pk.edu.pl

[2] Cracow University of Technology, Faculty of Mechanical Engineering
Poland, 31-864 Kraków, Al. Jana Pawła II 37
mariusz.krawczyk@mech.pk.edu.pl

[3] Cracow University of Technology, Faculty of Mechanical Engineering
Poland, 31-864 Kraków, Al. Jana Pawła II 37
anna.kielbus@mech.pk.edu.pl

Key words: Uncertainty Assessment, Fuzzy Logic, Probability, Prediction, DOE, bootstrap.

Abstract. This article reviews known approaches to determining the uncertainty of predictive models: probabilistic analytical, probabilistic simulation and fuzzy. The main elements determining the specificity of a given approach are shown. The advantages and disadvantages are compared. Finally, the application guidelines are listed.

1   INTRODUCTION

Modern civilization, by exploiting natural resources, has led biological and climatic ecosystems to the edge of catastrophic collapse. This path of civilization development is no longer possible, hence the implementation of methods and technologies of sustainable development is required. However, they require the use of new materials and processes of much greater complexity than those used so far. The development of new materials and technologies for their production is increasingly a planned design activity and less and less a random discovery.

In the field of materials science, this involves the intensive use of automated analytical lines that systematically search for possible combinations of the characteristics of raw materials and determine the properties of the obtained product materials, e.g. metallic glass [1, 2]. The bottom-up approach uses microstructure modeling and aggregation to examine the features and properties at the macroscale level e.g. bio-tissues [3, 4] or metallic foams [5, 6]. Due to the fact that it is impossible to predict in advance what macroscopic features will result from setting the microscale features, the creation of utility predictive models is most often performed using the inverse problem technique [7, 8] using data-driven [9-11] parametric [12], semiparametric [13, 14] or non-parametric models [14].

In the field of technological process engineering, analytical tools related to the automation of production lines and the presence of numerous measuring sensors, which is commonly known as Industry 4.0, are increasingly used [15, 16]. This leads to the creation of very large databases collecting huge amounts of information [17]. These datasets are characterized by both a large number of records and a large dimensionality, that is, the number of measured process characteristics recorded in these records. Internal feedback loops occurring in these processes cause their optimization and control to be highly nonlinear, and thus require the use of special techniques of building predictive models. This requires the use of either the well-known DOE techniques [18, 19], often embedded in more general methodological approaches, such as Six Sigma [20], or newer non-parametric machine learning methods [21, 22].

In both areas, in materials science and technological processes, it is most often necessary to reduce the dimensionality of the data and identify the dominant factors influencing the interesting features. This is typically done by factor analysis [23] or principal component analysis [24], which is a variation of linear discriminant analysis [25]. In more complex cases, a much more difficult nonlinear discriminant analysis [25] is used.

Merely obtaining a predictive model, whether parametric or nonparametric, is not enough. The predictive model positioning the mean value, but not defining the uncertainty of the result, has little industrial or research utility, hence it is very important to determine the uncertainty of the prediction.

In industrial and research practice, the most common is to determine the uncertainty of prediction by specifying the confidence interval with a predetermined confidence level [26], usually 95%. The prediction uncertainty formulated in this way is used quantitatively, although practitioners do not think about the very idea of this quantity at all: whether it is formulated in terms of the frequency [27], Bayesian [28] or propensity approach [29]. The formal mathematical apparatus is identical in all variants – Kolmogorov's formalism [30] – differences appear at the stage of formulating issues and interpreting the results. The most common approach is to use the Bayesian approach, even if it is not explicitly expressed, because the probabilities of single or low-frequency events are calculated, and the calculated value is treated as a subjective assessment of belief that the event will occur. From a purely practical point of view, uncertainty estimates are computed either from asymptotic analytical estimates or from numerical Monte Carlo simulations.

An alternative approach to uncertainty assessment and used much less frequently is the fuzzy assessment, the idea of which comes from L. Zadeh [31, 32]. This approach is based on an assessment of membership, which is largely subjectively assessed in a manner similar to the Bayesian belief. It is not, however, the field of Borel sets and the algebra of fuzzy operations is different, in addition with a rather arbitrary selection of pairs of functions that perform the sum and product of fuzzy assessments. Buckley [33] proposed a specific approach linking statistics to fuzzy membership: the fuzzy membership becomes the confidence level of the appropriate interval [34]. The fuzzy approach was intensively adapted to the needs of building regression models [35-37], recently [38] with specific intuitionistic approach [39]. Grzegorzewski worked intensively on the problem of real fuzzy numbers [40], especially in the trapezoidal approach [41], and then developed a complete approach to the issue of decision-making [42] building an ontology and taxonomy that allowed to consider various variants of the location of fuzzy evaluation elements.

2 COMPARISON OF THE APPROACHES

2.1 Probabilistic analytical-based uncertainty

The initial step is to identify the source data set, postulate the form of the predictive model, and then make an assumption about the characteristics of the random term [26]: its location in the model and the type of probability distribution. In addition, dataset preprocessing functions often appear, which aim either to obtain the desired statistical features e.g. normality, homogeneity, linearity or to meet the physical constraints by the predictive model: predicted values are non-negative or limited on both sides.

Due to such numerous assumptions, limitations and distortions, a simple composition is usually applied: a linear model (with respect to unknown parameters), a non-biased random term with the assumed normal distribution and the least squares criterion. Most of the uncertainty estimates, significance tests and identification methods are designed specifically for such a simple predictive model.

The advantages of these approaches are that they are well documented, widely available in the literature, and the procedures are incorporated into most popular computational programs and computational libraries.

A large disadvantage of these approaches is the numerous assumptions made, which are rarely taken into account and verified during application. Another rarely considered element of these approaches is the fact that most of them do not have exact mathematical forms and, contrary to popular belief, they are only rough asymptotic estimates.

2.2 Probabilistic simulation-based uncertainty

The first step is to identify the source dataset, postulate the form of the predictive model and locate the random term that meets the i.i.d. (independent and identically distributed) assumption [43-45]. The next step is to use the resampling procedure multiple times to obtain bootstrap samples to build a substitute for samples from a large population. This approach avoids making assumptions about a specific form of the probability distribution, and thus the obtained results are less biased.

This approach can be widely used, especially for large and huge datasets or in the absence of information or poor information on the type of distribution. This allows for the estimation of the uncertainty of both the parameters of the predictive model and the values it provides, both in the discrete factor approach [46] and the continuous response surface [47]. The simulation approach also allows for effective uncertainty assessment in the case of both the analysis of variance ANOVA [48, 49] and the principal component analysis PCA [50], although it requires a significant reconstruction of the calculation scheme.

The advantage of this approach is that there are no risky and questionable assumptions about the specific form of the random distribution. The adoption of the data-driven approach significantly improves the quality and accuracy of prediction.

There are two disadvantages to this approach. The first is the need to precisely locate the random term in the model structure and ensure that it meets the i.i.d. conditions. Failure to do so leads to worthless results. The second is the requirement to provide large computing power, both in terms of data processing speed and memory resources.

2.3 Fuzzy-based uncertainty

Contrary to the probabilistic approach, there is practically no distinction between the analytical and simulation approaches. The computational complexity and arbitrariness of the selection of T-S norms (product-sum operators) make analytical considerations very general, and any applications quickly come down to numerical simulations.

Grzegorzewski [42] proposed the directions of generalizations for the classical theory of statistical hypothesis verification. He defined three elements that can be treated in a fuzzy manner: data (e.g., three measurements of about 5, about 6, and between 9 and 10), hypotheses (e.g., a mean of about 10), and requirements (e.g., a significance level of about 0.05). This triplet may be mixed in any manner between strict and fuzzy description. In practice, due to both computational and purely conceptual difficulties, the most common are variants in which only one element of the triplet is fuzzy.

In the case of prediction, the problem comes down to determining a predictive model for a set of fuzzy data, and then determining the fuzzy uncertainty for the predicted values. The use of polynomial models identified on the basis of the least squares method is well recognized [35-37]. In recent years, fuzzy prediction using logistic regression, i.e. in relation to categorical values and latent explanatory variables, has also started to gain popularity [51, 52].

The advantage of this approach is the adaptation of imprecise and incomplete data descriptions [53]. This allows to avoid substantively questionable situations when poorly justified data imputation is made only to fill the required minimum input dataset of the analytical procedure.

The disadvantages of this approach are the need to very skilfully design the calculation scheme and the conceptually correct interpretation of the results.

12 CONCLUSIONS

The above considerations lead to the conclusion that there are three basic criteria for selecting a convenient approach for estimating prediction uncertainty: normality of distribution, completeness of data, precision of data description.

Imprecise data should be analyzed using the fuzzy approach. Precisely described and complete data with high confidence in meeting normality should be analyzed using a probabilistic analytical approach. For all other cases, the appropriate approach is to use a probabilistic numerical simulation.

The above considerations may be useful as guidelines for analyzes in scientific or industrial applications, including when examining the properties of the surface layer [54], corrosion resistance [55, 56] or analyzing incomplete or uncertain structural data [57, 58]. It may also be very useful in predicting the properties of living organisms [59, 60], for which the description is often incomplete, uncertain and dominated by individual traits. It is similar in the case of prediction of the occurrence of defects or failures [61, 62], when the history of events is uncertain and incomplete [63] or the physical model of the phenomena is only rough [64, 65] or disturbed by uncontrolled or weakly controlled factors [66].

REFERENCES

[1]   Samavatian, M., Gholamipour, R. and Samavatian, V. Discovery of novel quaternary bulk metallic glasses using a developed correlation-based neural network approach. *Computational Materials Science* (2021) **186**: art. 110025.

[2]   Wu, J.Q., Sun, Y.T., Wang, W.H. and Li, M.Z. Application of machine learning approach in disordered materials. *Scientia Sinica-Physica Mechanica & Astronomica* (2020) **50**(6): art. 067002.

[3]   Zhou, X.H., Zhang, E.H., Xia, M.G., Zhang, L., Tian, Z.Q., Liu, J.L. and Zhang, S.L. Structure-property relationships of cell clusters in biotissues: 2D analysis. *Physical Chemistry Chemical Physics* (2017) **19**(18): 11603-11611.

[4]   Zhuravkov, M., Drozd, L., Romanova, N. and Krupoderov, A. Mechanical-Mathematical Modelling of Biological Tissue Behaviour. *Analytic Methods in Interdisciplinary Applications* (2015) **116**: 153-181.

[5]   Wejrzanowski, T., Cwieka, K., Skibinski, J., Brynk, T., Ibrahim, S.H., Milewski, J. and Xing, W. Metallic foam supported electrodes for molten carbonate fuel cells. *Materials & Design* (2020) **193**: art. 108864.

[6]   Haj Ibrahim, S., Neumann, M., Klingner, F., Schmidt, V. and Wejrzanowski, T. Analysis of the 3D microstructure of tape-cast open-porous materials via a combination of experiments and modeling. *Materials & Design* (2017) **133**: 216-223.

[7]   Kuravi, R., Leichsenring, K., Trostorf, R., Morales-Orcajo, E., Böl, M. and Ehret, A.E. Predicting muscle tissue response from calibrated component models and histology-based finite element models. *Journal of the Mechanical Behavior of Biomedical Materials* (2021) **117**: art. 104375.

[8]   Hasan, M., Senthilnathan, A. and Acar, P. Machine learning reinforced multi-scale modeling of microstructures under uncertainties. *AIAA Scitech 2021 Forum* (2021): 1-13.

[9]   Panda, N., Osthus, D., Srinivasan, G., O'Malley, D., Chau, V., Oyen, D. and Godinez, H. Mesoscale informed parameter estimation through machine learning: A case-study in fracture modeling. *Journal of Computational Physics* (2020) **420**: art. 109719.

[10]   Korneev, S., Arunachalam, H., Onori, S. and Battiato, I. A Data-Driven Multiscale Framework to Estimate Effective Properties of Lithium-Ion Batteries from Microstructure Images. *Transport in Porous Media* (2020) **134**(1): 173-194.

[11]   Herriott, C. and Spear, A.D. Predicting microstructure-dependent mechanical properties in additively manufactured metals with machine- and deep-learning methods. *Computational Materials Science* (2020) **175**: art. 109599.

[12]   Heinz, S., *Mathematical Modeling*. Heidelberg: Springer (2011).

[13]   Horowitz, J.L., *Semiparametric Models*, in *Handbook of Computational Statistics*, J.E. Gentle, W.K. Hardle, and Y. Mori, Editors. 2012, Springer: Heidelberg. p. 597-617.

[14]   Härdle, W., Werwatz, A., Müller, M. and Sperlich, S., *Nonparametric and Semiparametric Models*. Heidelberg: Springer (2004).

[15]   Ingaldi, M. and Ulewicz, R. Problems with the implementation of industry 4.0 in enterprises from the SME sector. *Sustainability (Switzerland)* (2020) **12**(1): art. 217.

[16]   Karpisz, D. and Kiełbus, A. Selected problems of designing modern industrial databases. *MATEC Web of Conferences* (2018) **183**: art. 01017.

[17] Havard, V., Sahnoun, M., Bettayeb, B., Duval, F. and Baudry, D. Data architecture and model design for Industry 4.0 components integration in cyber-physical production systems. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* (2020): DOI: 10.1177/0954405420979463.

[18] Pietraszek, J., Radek, N. and Goroshko, A.V. Challenges for the DOE methodology related to the introduction of Industry 4.0. *Production Engineering Archives* (2020) **26**(4): 190-194.

[19] Pietraszek, J. and Goroshko, A. The heuristic approach to the selection of experimental design, model and valid pre-processing transformation of DoE outcome. *Advanced Materials Research* (2014) **874**: 145-149.

[20] Kregel, I., Stemann, D., Koch, J. and Coners, A. Process Mining for Six Sigma: Utilising Digital Traces. *Computers and Industrial Engineering* (2021) **153**: art. 107083.

[21] Pietraszek, J., Gądek-Moszczak, A. and Radek, N. The estimation of accuracy for the neural network approximation in the case of sintered metal properties. *Studies in Computational Intelligence* (2014) **513**: 125-134.

[22] Juan, Y., Dai, Y., Yang, Y. and Zhang, J. Accelerating materials discovery using machine learning. *Journal of Materials Science and Technology* (2021) **79**: 178-190.

[23] Bartholomew, D., Knott, M. and Moustaki, I., *Latent Variable Models and Factor Analysis*. Hoboken: Wiley (2011).

[24] Jolliffe, I.T., *Principal Component Analysis*. New York: Springer (2010).

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

[26] Montgomery, D.C., *Design and Analysis of Experiments. Tenth Edition.* Hoboken: Wiley (2020).

[27] Wakefield, J., *Bayesian and Frequentist Regression Methods*. New York: Springer Science + Business Media (2013).

[28] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B., *Bayesian Data Analysis. Third Edition.* Boca Raton: CRC Press (2013).

[29] Guo, S. and Fraser, M.W., *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks: SAGE (2015).

[30] Kolmogorov, A.N. and Morrison, N., *Foundations of the Theory of Probability: Second English Edition*. Mineola, New York: Dover Publications, Inc. (2018).

[31] Zadeh, L.A. Probability Measures of Fuzzy Events. *Journal of Mathematical Analysis and Applications* (1968) **23**(2): 421-&.

[32] Zadeh, L.A. Fuzzy Sets. *Information and Control* (1965) **8**(3): 338-&.

[33] Buckley, J.J., *Fuzzy statistics*. Heidelberg: Springer-Verlag (2004).

[34] Buckley, J.J., *Fuzzy probability and Statistics*. Heidelberg: Springer (2006).

[35] Bardossy, A. Fuzzy least squares regression: theory and applications. *Fuzzy Regression Analysis* (1992): 181-193.

[36] Buckley, J.J. and Feuring, T. Linear and non-linear fuzzy regression: Evolutionary algorithm solutions. *Fuzzy Sets Syst.* (2000) **112**(3): 381-394.

[37] D'Urso, P. and Gastaldi, T. An 'orderwise' polynomial regression procedure for fuzzy data. *Fuzzy Sets Syst.* (2002) **130**: 1-19.

[38] Pourahmad, S., Ayatollahi, S.M.T. and Taheri, S.M. Fuzzy Logistic Regression: A New Possibilistic Model and Its Application in Clinical Vague Status. *Iranian Journal of Fuzzy Systems* (2011) **8**(1): 1-17.

[39] Atanassov, K.T. Intuitionistic fuzzy sets. *Fuzzy Sets Syst.* (1986) **20**: 87-96.

[40] Grzegorzewski, P. Nearest interval approximation of a fuzzy number. *Fuzzy Sets Syst.* (2002) **130**: 321-330.

[41] Grzegorzewski, P. Trapezoidal approximation of fuzzy number. *Fuzzy Sets and Systems - IFSA 2003.* (2003): 237-244.

[42] Grzegorzewski, P., *Wspomaganie decyzji w warunkach niepewności. Metody statystyczne dla nieprecyzyjnych danych.* Warszawa: AOW EXIT (2006).

[43] Owen, A.B., *Empirical Likelihood.* Boca Raton: CRC Press (2001).

[44] Pawitan, Y., *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford: Clarendon Press (2001).

[45] Shao, J. and Tu, D., *The Jackknife and Bootstrap.* New York: Springer (1995).

[46] Pietraszek, J., Dwornicka, R. and Szczotok, A. The bootstrap approach to the statistical significance of parameters in the fixed effects model. *ECCOMAS Congress 2016 - Proceedings of the 7th European Congress on Computational Methods in Applied Sciences and Engineering* (2016) **3**: 6061-6068.

[47] Pietraszek, J. and Wojnar, L. The bootstrap approach to the statistical significance of parameters in RSM model. *ECCOMAS Congress 2016 - Proceedings of the 7th European Congress on Computational Methods in Applied Sciences and Engineering* (2016) **1**: 2003-2009.

[48] Gądek-Moszczak, A., Pietraszek, J., Jasiewicz, B., Sikorska, S. and Wojnar, L. The bootstrap approach to the comparison of two methods applied to the evaluation of the growth index in the analysis of the digital X-ray image of a bone regenerate. *Studies in Computational Intelligence* (2015) **572**: 127-136.

[49] Pietraszek, J., Szczotok, A., Kołomycki, M., Radek, N. and Kozień, E. Non-parametric assessment of the uncertainty in the analysis of the airfoil blade traces. *METAL 2017 - 26th International Conference on Metallurgy and Materials* (2017): 1412-1418.

[50] Pietraszek, J. and Skrzypczak-Pietraszek, E. The uncertainty and robustness of the principal component analysis as a tool for the dimensionality reduction. (2015) **235**: 1-8.

[51] Pourahmad, S., Ayatollahi, S.M.T., Taheri, S.M. and Agahi, Z.H. Fuzzy Logistic Regression Based on the Least Squares Approach with Application in Clinical Studies. *Computer & Mathematics with Applications* (2011) **62**(9): 3353-3365.

[52] Yang, M.S. and Chen, H.M. Fuzzy Class Logistic Regression Analysis. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems* (2004) **12**(6): 761-780.

[53] Pietraszek, J., Sobczyk, A., Skrzypczak-Pietraszek, E. and Kołomycki, M. The fuzzy interpretation of the statistical test for irregular data. *Technical Transactions* (2016) **113**: 119-126.

[54] Dwornicka, R., Radek, N., Krawczyk, M., Osocha, P. and Pobędza, J. The laser textured surfaces of the silicon carbide analyzed with the bootstrapped tribology

model. *METAL 2017 - 26th International Conference on Metallurgy and Materials, Conference Proceedings* (2017): 1252-1257.

[55] Pliszka, I. and Radek, N. Corrosion Resistance of WC-Cu Coatings Produced by Electrospark Deposition. *Procedia Engineering* (2017) **192**: 707-712.

[56] Lipinski, T. and Karpisz, D. Effect of animal slurry on carbon structural s235jr steel at 303 k. *Engineering for Rural Development* (2020) **19**: 1482-1487.

[57] Korzekwa, J., Gądek-Moszczak, A. and Zubko, M. Influence of the Size of Nanoparticles on the Microstructure of Oxide Coatings. *Materials Science* (2018) **53**(5): 709-716.

[58] Bochenek, D., Niemiec, P., Korzekwa, J., Durtka, B. and Stokłosa, Z. Microstructure and properties of the ferroelectric-ferromagnetic PLZT-ferrite composites. *Symmetry* (2018) **10**(3).

[59] Skrzypczak-Pietraszek, E., Piska, K. and Pietraszek, J. Enhanced production of the pharmaceutically important polyphenolic compounds in Vitex agnus castus L. shoot cultures by precursor feeding strategy. *Engineering in Life Sciences* (2018) **18**(5): 287-297.

[60] Wojnar, L., Gadek-Moszczak, A. and Pietraszek, J. On the role of histomorphometric (stereological) microstructure parameters inthe prediction of vertebrae compression strength. *Image Analysis and Stereology* (2019) **38**(1): 63-73.

[61] Pacana, A., Czerwińska, K. and Dwornicka, R. Analysis of non-compliance for the cast of the industrial robot basis. *METAL 2019 - 28th International Conference on Metallurgy and Materials, Conference Proceedings* (2019): 644-650.

[62] Trzewiczek, K., Szczotok, A. and Gadek-Moszczak, A. Evaluation of the state for the material of the live steam superheater pipe coils of V degree. *Advanced Materials Research* (2014) **874**: 35-42.

[63] Karpisz, D. Design of manufacturing databases. *Technical Transactions* (2016) **113**: 73-77.

[64] Danielewski, H. Laser welding of pipe stubs made from super 304 steel. Numerical simulation and weld properties. *Technical Transactions* (2019) **116**: 167-176.

[65] Radek, N., Kurp, P. and Pietraszek, J. Laser forming of steel tubes. *Technical Transactions* (2019) **116**: 223-229.

[66] Gazda, A., Pacana, A. and Malindžák, D. Study on improving the quality of stretch film by Taguchi method. *Przemysl Chemiczny* (2013) **92**(6): 980-982.