AIP Conference Proceedings

Study of the tensile strength of alloy steels using polynomial regression

Cite as: AIP Conference Proceedings **2522**, 050005 (2022); https://doi.org/10.1063/5.0101206
Published Online: 26 September 2022

S. Gocheva-Ilieva and G. Dobrev

View Online Export Citation

AIP Conference Proceedings **2522**, 050005 (2022); https://doi.org/10.1063/5.0101206 **2522**, 050005

Study of the Tensile Strength of Alloy Steels Using Polynomial Regression

S. Gocheva-Ilieva[1,a)] and G. Dobrev[2,b)]

[1] *Paisii Hilendarski University of Plovdiv, 24 Tsar Asen str., 4000 Plovdiv, Bulgaria*
[2] *Technical University Sofia –Plovdiv Branch, 25 Tsanko Dyustabanov str., 4000 Plovdiv, Bulgaria*

[a)]Corresponding author: snow@uni-plovdiv.bg
[b)] dobrev.1975@abv.bg

Abstract. The objective of this study is to identify the influence of the chemical composition of alloyed steel on the tensile strength through predictive multiple regression models of the first and second degrees. Data on the percentage content of nine chemical compounds in alloy steel are used as independent variables: C, Cr, Mn, Mo, Ni, P, S, Si, Al, and the product's diameter. In order to accurately perform multivariate linear regression, all variables undergo Yeo-Johnson transformation in advance to achieve normal or near-normal distribution. The obtained linear regression models fit the measured values for tensile strength with 76% and the models with predictors up to second order – with 97.6%. The alloy compounds with the strongest influence on tensile strength are identified.

INTRODUCTION

The mechanical properties of metals have always been subject to experimental and theoretical research. They determine the ability of metals to resist various forms of strain and destruction during use. They can be subjected to various types of loads – tension, pressure, bending, torsion, etc. Often these loads can be combined, for example tension and bending, tension and torsion, etc.

Various experimental research is carried out to expand the scope of application of steels. This is essential for the development of new types of steel and products made from them [1,2]. Along with technical measurements and experiments, mathematical and statistical methods are also used in research [3]. The goal is to investigate the properties of known samples of steels and also to predict the behavior of new ones under the preset operating requirements.

There are numerous empirical results for the considered topic in scientific literature. Various statistical methods are used to study the properties of alloy steels through different data. The methods can be classified in two general groups – multivariate statistical analysis and machine learning. The first group encompasses the classical approaches such as correlation and regression analysis, factor analysis, Principal component analysis (PCA), ANOVA, MANOVA, etc. The second group focuses on the use of machine learning and data mining methods, such as Artificial Neural Networks (ANN), Support vector machine regression, Random Forest regression, fuzzy functions, and others.

We will consider in more detail some representative publications from the first group. The Weibull distribution method is used in [4] to study the tensile strength of alloy steels at room and elevated temperatures. The authors of [5] build polynomial type regression models to identify statistically significant variables for predicting experimental measurements of the tensile strength and yield strength at room and elevated temperatures of stainless steel 17-7PH with modified chemical composition. In [6], the conditions needed to reduce the time and costs related to the development and selection of new metal alloys are determined. The study is based on a multiple linear regression model for predicting tensile strength of low alloy steel depending on its chemical composition (C, Mn, Cr, Ni, Mo, Cu, N, V), plate thickness, type of solution used to process it and aging temperatures. PCA is used in [7] for optimization of tribological characteristics of aluminum metal matrix composites for achieving better wear

Application of Mathematics in Technical and Natural Sciences
AIP Conf. Proc. 2522, 050005-1–050005-8; https://doi.org/10.1063/5.0101206
Published by AIP Publishing. 978-0-7354-4361-7/$30.00

properties. In the recent paper [8], the influence of 9 chemical elements (C, Cr, Mn, Mo, Ni, P, S, Si, Al) on the tensile strength of alloy steels is classified with the aid of factor and cluster analyses. A detailed review of the results obtained through classical and machine learning statistical techniques in the field of stainless steels can be found for example in [3,9,10] among others.

The objective of this study is to examine the relationship between the mechanical property tensile strength and the content of various metals in alloy steel and the diameter of the measured test sample. In order to establish the statistical relationship, regression polynomial models of first and second degree are built and studied. The models also enable forecasting the experiment and identifying boundaries for the expected tensile strength value of alloy steels.

Statistical studies are carried out in the environment of IBM SPSS Statistics [11].

PROBLEM SETUP AND METHODS USED

Our main task is to show how with available measurements for a given property of a set of alloy steels, adequate statistical models can be built both to study the influence of individual independent variables on the property and to predict future experiments. The property of tensile strength of alloyed steels and their chemical composition are chosen as an empirical example.

Multiple Linear Regression

The use of standard linear regression models to study the properties of steels as a function of their chemical composition is not new. Usually, the values of the studied property   of steel are represented as a multiple linear regression (MLR) model of the type:

$$Y = a_0 + \sum_{k=1}^{p} a_k X_k + \varepsilon, \qquad \varepsilon \in N(0, s^2) \tag{1}$$

where X_k is the chemical element in the composition of the of the steel product or other predictor, a_k, $k = 0,1,...,p$ – regression coefficients, ε – model error. For the error, it needs to be established that it is white noise, i.e. normally distributed with a zero mean and variance s^2 .

To obtain an adequate MLR of the type (1), it is necessary to take into account the performance of statistical assumptions such as the presence of a linear type relationship between the independent variables X_k and the dependent variable Y , normality of variables, *etc.* [12]. In many cases, there is a multicollinearity effect between the independent variables that reduces significantly the accuracy of the applied algorithms of the regression method, which needs to be avoided.

Polynomial Regression Models

It is well known that the influence of alloying elements on the mechanical properties of steels and in particular on tensile strength is characterized by both linear and non-linear complexity. To find adequate and more accurate models, in our study we build and consider polynomial models where the independent variables participate up to the second degree.

In this paper, our objective is to demonstrate that the accurate application of polynomial type regression models makes it possible to achieve higher predictive power which is in no way inferior to modern machine learning techniques. At the same time, we will note the main advantages of this approach, such as the simple and straightforward interpretation of the results of polynomial models and the low price of their computer implementation including the use of memory and processor time.

DATA AND INITIAL DATA PROCESSING

Data from [8] are used in this study, including 60 different types of alloy steels, selected from publications [13,14]. Descriptive statistics of the initial variables are given in Table 1. Here σ, MPa denotes the dependent variable tensile strength. The percentage content of the nine chemical elements in alloy steels and the diameter *d*, mm of the measured test sample are considered as independent variables.

TABLE 1. Descriptive statistics of the initial measured variables.

Variable	N valid	Mean	Median	Std. Deviation	Variance	Minimum	Maximum
σ, MPa	60	738.133	681.5	279.69	78224.66	370	1375
C, %	60	0.22	0.175	0.108	0.012	0.03	0.47
Cr, %	60	2.614	0.5	5.705	32.545	0	25
Mn, %	60	0.858	0.6	0.473	0.223	0.45	2
Mo, %	60	0.229	0	0.41	0.168	0	2.4
Ni, %	60	1.643	0	2.743	7.523	0	9.25
P, %	60	0.081	0.04	0.1	0.01	0.02	0.35
S, %	60	0.138	0.04	0.349	0.122	0	2.35
Si, %	60	0.42	0.375	0.238	0.057	0.1	1
Al, %	60	0.009	0	0.049	0.002	0	0.375
d, mm	60	94.7	50	112.75	12712.25	3	500

None of the eleven variables, presented in Table 1, have normal distribution. Also, the ten independent variables clearly demonstrate multicollinearity, which is established in [8]. This directly impedes the application of MLR. To improve the distribution and stabilize the variance, each variable undergoes Yeo-Johnson transformation beforehand, which in this case is used only for the negative values [15]:

$$ty = \begin{cases} \dfrac{(x+1)^{\lambda} - 1}{\lambda}, & x \geq 0,\ \lambda \neq 0 \\ \ln(x+1), & x \geq 0,\ \lambda = 0 \end{cases}, \quad \lambda \in [-2,\ 2]. \tag{2}$$

In (2), x is the input variable, ty is the transformed one, and the parameter λ is found so that the distribution is normal or as close to normal as possible. For the variables in Table 1, the following coefficients λ are determined and given in Table 2.

In order to study the possible multicollinearity among the transformed variables, the correlation matrix is calculated using Pearson bicorrelation coefficients. This does not lead to significantly high absolute values of the coefficients. The highest is between *tP* and *tS*, equal to 0.712. The determinant of the correlation matrix is 0.017, which is significantly different from zero. Also, the highest variance inflation index (VIF) is equal to 4.499, so that for all variables VIF <10. In accordance with [16,14)], we can conclude that there is no multicollinearity effect for the transformed independent variables, listed in Table 2.

TABLE 2. The values of the parameter λ when the variables are transformed.

Initial variable	Transformed variable	Value of λ
σ	*tσ*	-0.3
C	*tC*	-1
Cr	*tCr*	-0.8
Mn	*tMn*	-2
Mo	*tMo*	-2
Ni	*tNi*	-0.8
P	*tP*	-2
S	*tS*	-2
Si	*tSi*	-2
Al	*tAl*	-2
d	*td*	-0.1

MULTIPLE LINEAR REGRESSION MODEL OF TENSILE STRENGTH

The results from the previous section 3 enable the correct performance of multiple linear regression (MLR) using the transformed variables from Table 2. We apply the backward regression method. The resulting linear regression equation has the following form:

$$t\sigma_1 = 2.701 + 0.417\,tC + 0.171\,tMn + 0.180\,tMo + 0.077\,tNi + 0.524\,tP - 0.097\,tSi\,. \tag{3}$$

It includes only 6 of the 10 independent variables. The regression coefficients of the remaining four *tCr, tS, tAl* and *td* are statistically insignificant at a standard level of significance $\alpha = 0.05$. Equation (3) in standardized form is written out as

$$zt\sigma_1 = 0.558\,tC + 0.202\,tMn + 0.482\,tMo + 0.608\,tNi + 0.320\,tP - 0.142\,tSi\,. \tag{4}$$

From (4), it can be concluded that the values of tensile strength for the studied sample depend to a large extent on the contents of *Ni*, *C* and *Mo*. The weakest participation is that of the element *Si*, and negative at that, i.e. higher *Si* content reduces σ. It needs to be noted, that in this case only a linear relationship is found.

In order to compare the quality of the obtained linear model, we perform retransformation of the predicted non-standardized values in (3). The obtained model we denote with $\hat{\sigma}_1$. Its statistics are given on the first row of Table 3. The resulting coefficient of determination $R^2 = 0.762$ shows that the model $\hat{\sigma}_1$. accounts for the data sample at 76%. Its Root Mean Squared Error (RMSE) is 136.673. The maximum relative error is very big - $\Delta_1 = 30\%$.

The quality of fit is shown in Figure 1 sorting the data on tensile strength.

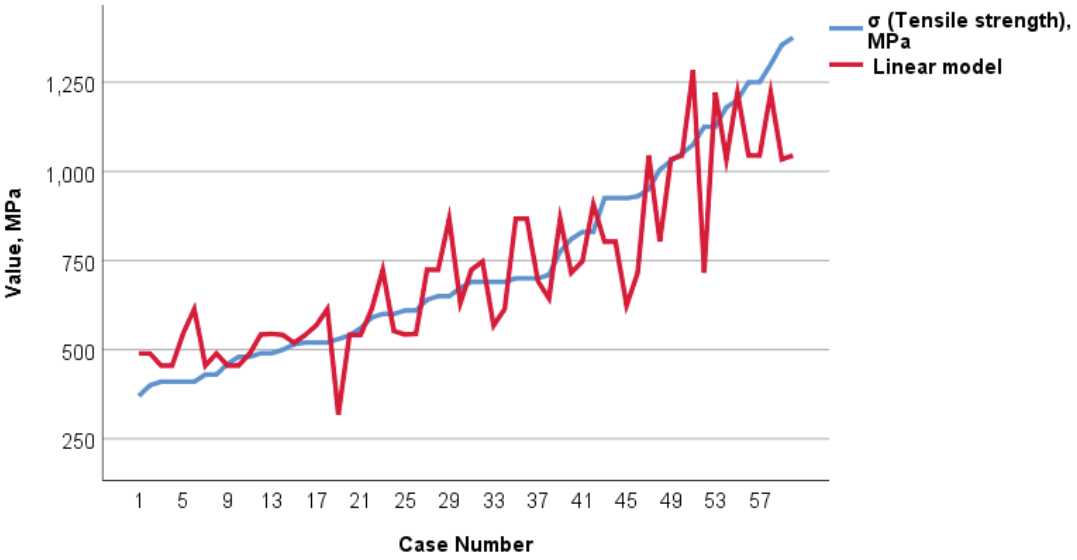

FIGURE 1. Line plots of measured values of tensile strength and their predictions from the linear regression model $\hat{\sigma}_1$

A scatter-plot of measured values versus values predicted by the linear model is presented in Figure 2. Some relatively big deviations are observed.

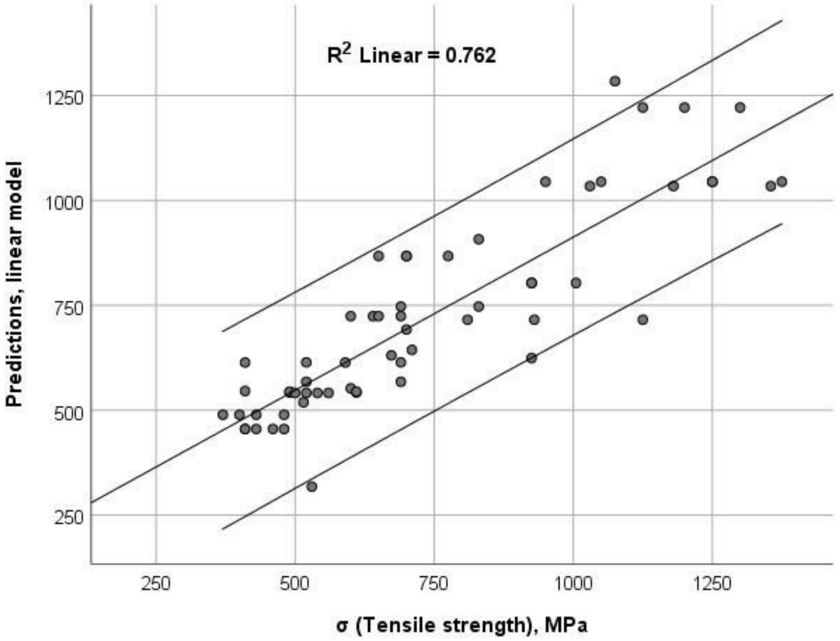

FIGURE 2. Measured versus predicted values of tensile strength using the obtained linear model $\hat{\sigma}_1$ with 5% confidence interval

The error analysis of the linear model shows that they have a normal distribution. The results of the statistical normality tests are as follows: for the Kolmogorov-Smirnov test, the significance is 0.052 and the significance of the Shapiro-Wilk test is 0.260. We can assume that error of model $\hat{\sigma}_1$ is a white noise.

SECOND DEGREE POLYNOMIAL MODEL

The use of linear regression models to present various physical processes and properties of compounds does not account for the non-linear nature of processes in physics. The world we live in is non-linear. Chemical elements influence the tensile strength not just on their own but in combination with other chemical elements. This influence cannot be assessed with linear interactions alone. In this case, we will build a second degree non-linear model of $t\sigma$, which includes not only the 10 separate independent variables form Table 2 but also their products in pairs

$$tX_k.tX_j, \quad k \le j, \quad k, j = 1, 2, ..., 10 . \tag{5}$$

To this end we calculate 55 new variables. We apply backward linear regression with a total of 65 predictors. Of all independent variables used, at each step statistically insignificant ones are dropped. The final model is clarified after 13 steps. Thus, out of all 65 input predictors, there are only 23 statistically significant left. The obtained regression model is denoted with $t\sigma_2$, and the retransformed one – with $\hat{\sigma}_2$. The obtained non-standardized equation has the following form

$$\begin{aligned} t\sigma_2 = 2.869 &+ 0.455tCr - 0.083td - 4.304tCtC - 1.866tC.tCr + 11.101tC.tMn - 5.757tC.tSi \\ &+ 9.763tC.tAl - 0.086tCr.tCr + 0.177tCr.tMo + 0.524tCr.tNi - 0.043tCr.td - 2.817tMn.tMn \\ &- 1.565tMn.tMo + 0.147tMn.td - 0.363tMo.tNi + 0.145tMo.td + 0.137tNi.tNi - 1.657tNi.tP \\ &+ 0.289tNi.tSi - 2.809tNi.tAl - 0.013tNi.td - 1105.547tS.tAl + 0.139tSi.td \end{aligned} . \tag{6}$$

The respective standardized regression equation of the second degree polynomial model is:

$$\begin{aligned} t\sigma_2 = {} & 3.606tCr - 1.849td - 2.301tC.tC - 2.367tC.tCr + 5.370tC.tMn - 1.680tC.tSi \\ & + 0.262tC.tAl - 0.682tCr.tCr + 0.399tCr.tMo + 3.639tCr.tNi - 1.794tCr.td - 2.338tMn.tMn \\ & - 1.229tMn.tMo + 1.340tMn.td - 0.587tMo.tNi + 2.186tMo.td + 1.040tNi.tNi - 0.809tNi.tP \quad . \\ & + 0.686tNi.tSi - 1.712tNi.tAl - 0.524tNi.td - 2.576tSt.Al + 1.134tSi.td \end{aligned} \tag{7}$$

In equation (6), respectively (7), only two significant linear members are left – tCr and td. The second degree members contain all variables to a different extent. The largest number of interactions are those with *tNi* (7 times), *tC* and *tCr* – 5 times each, *tMo* and *td* – 4 times each, etc. After retransforming values of $t\sigma_2$, we obtain the final model $\hat{\sigma}_2$. The assessment of its summary statistics are given on the last row of Table 3. The goodness of fit measures are very high: $R^2 = 0.976$ and $\text{RMSE} = 42.737$. The relative error is within $\Delta_2 = 9\%$.

TABLE 3. Summary statistics of the obtained regression polynomial models.

Model	R	R^2	R^2 **Adjusted**	**Std. Error of the Estimate**	**RMSE**	**Maximum Relative Error**	**Sig. (p value)**
$\hat{\sigma}_1$	0.873	0.762	0.758	115.0997	136.673	$\Delta_1 = 30\%$	0.000
$\hat{\sigma}_2$	0.988	0.976	0.976	43.0212	42.737	$\Delta_2 = 8.8\%$	0.000

The analysis of the errors in the second degree non-linear model shows that they have normal distribution. The respective statistical normality tests are as follows: significance for the Kolmogorov-Smirnov test 0.200 and the significance of the Shapiro-Wilk test 0.203. We can conclude that the main requirements for the validity of the obtained regression model (6) - (7) are also met.

The comparison between the measured values for tensile strength and those predicted by the second degree polynomial model $\hat{\sigma}_2$ is illustrated in Figure 3. There is significant improvement compared to the results with the linear model of Figure 1. The next figure 4 shows the scatter-plot of the comparison with a 5% confidence interval. The quality of fit for the highest and the lowest values of tensile strength is significantly improved compared to Figure 2.

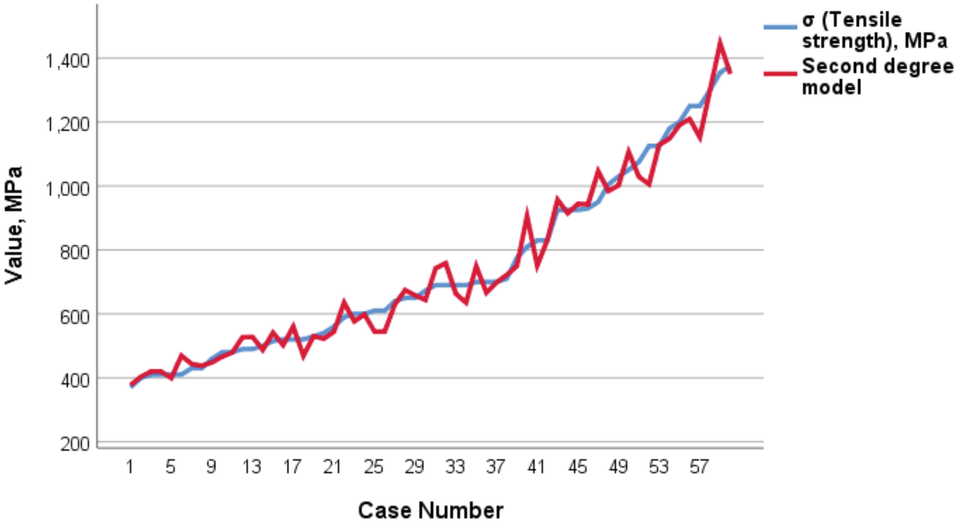

FIGURE 3. Line plots of measured values of tensile strength with their predictions from the second degree regression model $\hat{\sigma}_2$

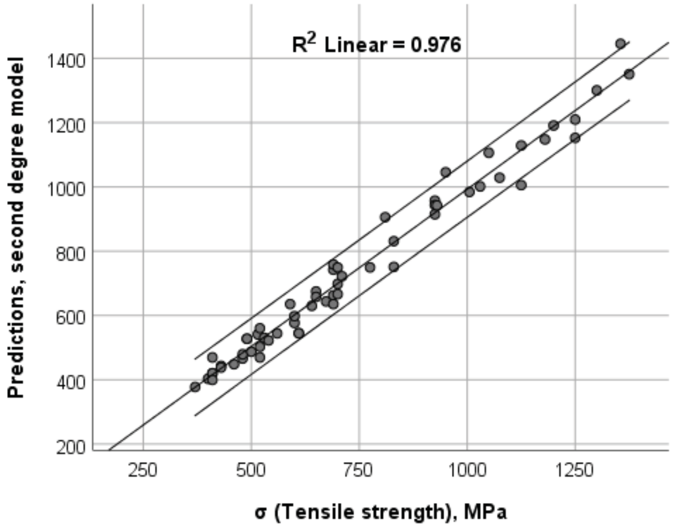

FIGURE 4. Measured versus predicted values of tensile strength using the obtained second degree model $\hat{\sigma}_2$ with 5% confidence interval

CONCLUSION

Two MLR models – linear and with the second degree of the predictors – are built to model and predict tensile strength – one of the main mechanical properties of alloy steels. It is shown that using a suitable transformation of the variables and accurate application of the MLR method, it is possible to obtain high performance models, reaching fit of the values predicted by the model and the measured values of tensile strength of up to 97.6%.

The developed regression models describe the tensile strength of alloy wheels in a different way. It is found that non-linear models provide a broader and more comprehensive interpretation of the influence of chemical elements contained in steel products. Regression equations indicate that the chemical elements affect significantly the tensile strength when they interact together in a complex pattern.

ACKNOWLEDGEMENTS

This study is supported by the Grant No BG05M2OP001-1.001-0003-C01, financed by the Science and Education for Smart Growth Operational Program (2014-2020), co-financed by the European Union through the European structural and Investment funds. The second author is grateful to the Research and Development Sector at the Technical University of Sofia, Bulgaria.

REFERENCES

1. D. T. Llewellyn and R. C. Hudd, *Steels: Metallurgy and Applications* (Butterworth-Heinemann, Oxford, 1998).
2. E. P. DeGarmo, J. T. Black, and R. A. Kohser, *Materials and Processes in Manufacturing*, 9th edn (Wiley, Hoboken, 2003).
3. K. H. Lo, C. H. She, and J. K. L. Lai, Recent developments in stainless steels, Materials Science and Engineering R **65**(4-6), 39–104 (2009), DOI: 10.1016/j.mser.2009.03.001.
4. P. SungHo, P. NoSeok, and K. JaeHoon, "A statistical study on tensile characteristics of stainless steel at elevated temperatures," in *Proceedings of 15th International Conference on the Strength of Materials (ICSMA-15)*, IOP Publishing Journal of Physics: Conference Series 240, 012083 (IOP Publishing Ltd., Bristol, UK, 2010), pp. 1–4, DOI:10.1088/1742-6596/240/1/012083.
5. B. Fakić, D. Ćubela, A. Burić, and E. Horoz, "Regression analysis of tensile strength testing results steel 17-7 PH with modified chemical composition," in *Proceedings of 14th International conference on Accomplishment in Mechanical and Industrial Engineering-DEMI 2019*, Banja Luka, 24-25 May 2019, pp. 691–697 (2019).

6. A. Golodnikov, Y. Macheret, A. A. Trindade, S. Uryasev, and G. Zrazhevsky, Statistical modelling of composition and processing parameters for alloy development, IOP Publishing Modelling Simul. Mater. Sci. Eng. **13**, 633–644 (2005), DOI:10.1088/0965-0393/13/4/013.
7. R. Siriyala, A. G. Krishna, P. R. M. Raju and M. Duraiselvam, Multi-response optimization of tribological characteristics of aluminum MMCs using PCA, Multidiscipline Modeling in Materials and Structures **10**(2), 276–287 (2014), DOI: 10.1108/MMMS-06-2013-0045.
8. G. L. Dobrev and I. P. Iliev, "Classification analysis of tensile strength of alloyed steels," in *IOP Conference Series: Materials Science and Engineering*, 878(1) (IOP Publishing Ltd., Bristol, UK, 2020), art. 012066, DOI: 10.1088/1757-899X/878/1/012066.
9. S. Datta and P. Chattopadhyay, Soft computing techniques in advancement of structural metals, International Materials Reviews **58**(8), 475–504 (2013), DOI: 10.1179/1743280413Y.0000000021.
10. M. J. Faizabadi, G. Khalaj, H. Pouraliakbar, and M. R. Jandaghi, Predictions of toughness and hardness by using chemical composition and tensile properties in microalloyed line pipe steels, Neural Computing and Applications **25**(7-8) 1993–1999 (2014), DOI:10.1007/s00521-014-1687-9.
11. IBM SPSS Statistics, https://www.ibm.com/analytics/spss-statistics-software
12. A. J. Izenman, *Modern Multivariate Statistical Techniques*. *Regression, Classification, and Manifold Learning* (Springer, New York, 2008).
13. J. E. Bringas (ed), *Handbook of Comparative World Steel Standards DS67C*, 4th edn (ASTM International, West Conshohocken, PA, 2007).
14. *Atlas Specialty Metals. Technical Handbook of Bar Products* (Atlas, Mowbray, 2005), http://www.atlassteels.com.au/documents/Atlas%20Engineering%20Bar%20Handbook%20rev%20Jan%202005-Oct%202011.pdf.
15. I. K. Yeo and R. A. Johnson, A new family of power transformations to improve normality or symmetry, Biometrika **87**(4), 954–959 (2000), DOI:10.1093/biomet/87.4.954.
16. E. R. Mansfield and B. P. Helms, Detecting multicollinearity, The American Statistician **36**(3) 158–160 (1982), DOI:10.2307/2683167.