Azerbaijan Focus Journal Of International Affairs, 10 (1) (2014) 01-10

www.jia.ueuo.com

COMPARISON OF LINEAR AND NONLINEAR STATISTICS METHODS APPLIED IN INDUSTRIAL PROCESS MODELING PROCEDURE

Predrag Đorđević, Ivan Mihajlović* and Živan Živković
University of Belgrade, Technical faculty in Bor
Vojske Jugoslavije 12, 19210 Bor, Serbia

Abstract

This paper presents the comparison of Multiple Linear Regression Analysis (MLRA) and Artificial Neural Networks (ANN) as the statistical analysis tools. Most influential statistical parameters for choosing right modeling tool are evaluated in this investigation. Investigation was performed on real statistical data set obtained after measurements of the process parameters underindustrial conditions.

*Keywords:*MLRA, ANN, statistical modeling

1. INTRODUCTION

The main objective of this study was to investigate applicability and constrains of the linear (MLRA) and nonlinear (ANN) methods of statistical analysis. Both approaches were used for modeling the same data set obtained during experiment facilitated under the industrial conditions.

The technological process that was the target of this investigation was aluminate solution decomposition, as the part of the Bayer alumina production process. However, technological process in question is not that much of interest in this paper. This paper is focused on the methodological process of modelling. Same methodological approach could be used for many other industrial processes.

In the process which was modeled in these investigations, following input variables were considered: concentration of the Na_2O (caustic); caustic ratio and crystallization ratio; starting temperature; final temperature; average diameter of crystallization seed and duration of decomposition process. Only one output variable was correlated above defined inputs of the process. Output variable considered was the degree of the aluminate solution decomposition.

2. MODELING PROCEDURE

In the contemporary systems theory literature, two main modeling approaches could be distinguished:

M1 – modeling procedure based on the system of differential equations resulting from cognizance of the systems structure and interdependence among the elements of the system (Weir, 1991; Brown, 2007; Dragićević and Bojić, 2009).

M2 – modeling procedure based on experimentally obtained data on the output of the system, resulting after introduction limited set of selected input parameters of the process (Taylor et al., 2003; Giraldo – Zuniga et al., 2006).

In this paper modeling approach M2 was selected. M2 approach could further be divided on analytical and statistical methods. For modeling described in this paper we used statistical methods, both linear (MLRA) and nonlinear (ANNs).

3. STARTING DATA SET

For modelling the process of aluminate solutions decomposition the data from thefactory Birač, Zvornik (Bosnia and Hercegovina), were used. The data was collected during the years 2008 - 2009, by measuring input and output process parameters, under stable operation mode of the production line. Total number of 500 data sets was collected this way, comprising:

a) Input parameters of the process: Na_2O (caustic) content in the solution (g/dm^3) – X_1; caustic ratio (α_k) of the solution – X_2; crystallization ratio – X_3; starting temperature of the solution (^oC) – X_4; final temperature of the solution (^oC) – X_5; average diameter of the crystallization seed (μm) – X_6; and the duration of the crystallization process (h) – X_7.

b) Output parameter of the process: degree of decomposition of the solution (%) – Y.

Table 1. Values of the input (Xi) and the output (Y) variables of the process of industrial sodium aluminate solution decomposition – descriptive statistics of 500 data sets

	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
				Statistic	Std. Error		
X_1	12.3	144.0	156.3	150.944	0.0762	1.7030	2.900
X_2	0.2	1.5	1.7	1.530	0.0015	0.0329	0.001
X_3	3.4	1.3	4.7	2.285	0.0296	0.6617	0.438
X_4	11.0	58.0	69.0	64.656	0.0536	1.1988	1.437
X_5	22.2	36.3	58.5	50.582	0.1839	4.1121	16.909
X_6	37.7	87.2	124.9	106.473	0.3806	8.5098	72.416
X_7	76.0	49.0	125.0	77.080	0.6236	13.9437	194.426
Y	24.3	32.3	56.6	46.658	0.1241	2.7747	7.699

4. RESULTS AND DISSCUSION

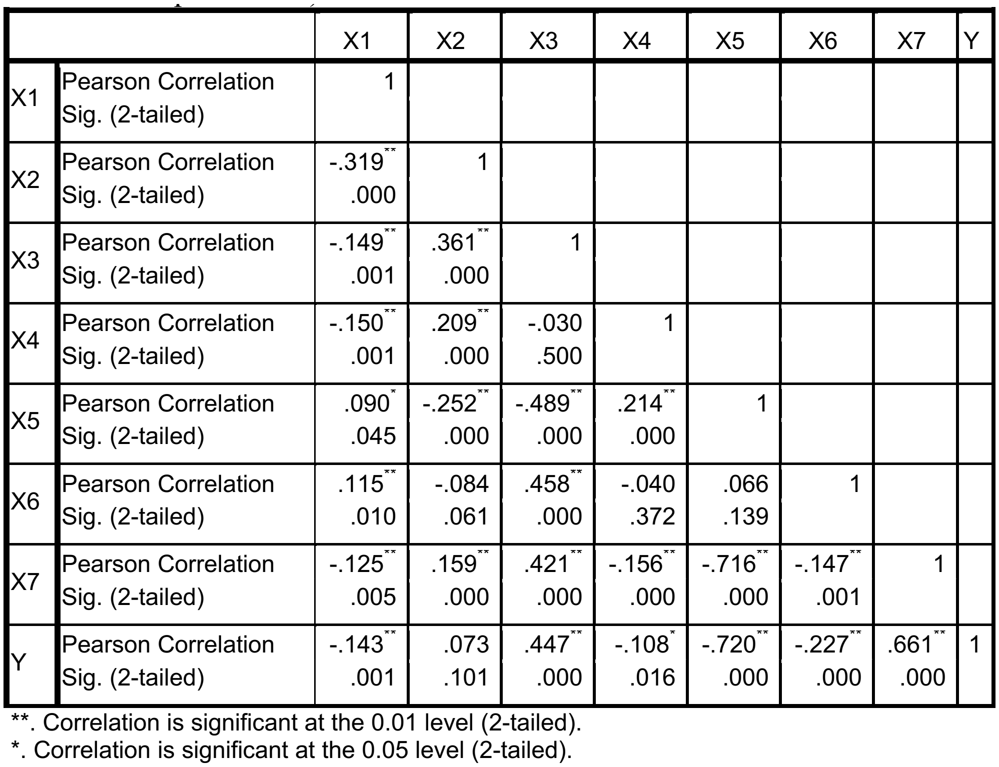
Values of the measured input parameters of the technological process (X_1 – X_7) and the process quality indicator – output of the process (Y) in the form of descriptive statistics results, are presented in table 1.

It can be noticed that variable X_2 has quite small variance (see Table 1). Nevertheless, this variable presents the caustic ratio of the solution and it is one of the most important parameters of the Bayer process. This way it shouldn't be omitted from subsequent analyse. Even small decrease of this variable leads to considerable increase of rate and the degree of decomposition of the solution. For example, if X_2 is changed from 1.7 to 1.5, the degree of decomposition (Y) will increase from 51% to 55%, with all other input parameters kept constant.

For definition of the correlation dependence in the form: output of the process (Y) = f input of the process (X_1 – X_7), bivariate correlation analysis was performed. As the result of this analysis Pearson Correlation (PC) coefficients with responding statistical significance were calculated, Table 2.

To finally define the dependence of the output parameter as the function of the input

Table 2. Correlation matrix for the input (X_1 – X_7) and the output (Y) variables of the industrial sodium aluminate solution dissociation process (number of data points for each variable is equal to 500)



		X1	X2	X3	X4	X5	X6	X7	Y
X1	Pearson Correlation Sig. (2-tailed)	1							
X2	Pearson Correlation Sig. (2-tailed)	-.319** .000	1						
X3	Pearson Correlation Sig. (2-tailed)	-.149** .001	.361** .000	1					
X4	Pearson Correlation Sig. (2-tailed)	-.150** .001	.209** .000	-.030 .500	1				
X5	Pearson Correlation Sig. (2-tailed)	.090* .045	-.252** .000	-.489** .000	.214** .000	1			
X6	Pearson Correlation Sig. (2-tailed)	.115** .010	-.084 .061	.458** .000	-.040 .372	.066 .139	1		
X7	Pearson Correlation Sig. (2-tailed)	-.125** .005	.159** .000	.421** .000	-.156** .000	-.716** .000	-.147** .001	1	
Y	Pearson Correlation Sig. (2-tailed)	-.143** .001	.073 .101	.447** .000	-.108* .016	-.720** .000	-.227** .000	.661** .000	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

parameters, using the multiple linear regression analysis (MLRA) with acceptable level of fitting (strong correlation), it is necessary that the value of PC is near 0.5 with statistical significance ($p \leq 0.05$) (Moroney, 1998; Živković et al., 2009a; Živković et al., 2009b). Analysis of the data presented in the Table 2 reveals that this constraint is attained in following cases: $Y - X_3$: PC = 0.447 (p = 0.000); $Y - X_5$: PC = - 0.720 (p = 0.000); $Y - X_7$: PC = 0.661 (p = 0.000). This was also the case for the following interdependence between the predictors of the process: $X_3 - X_5$: PC = -0.489 i p = 0.000; X_7 - X_5: PC = -0.716 i p = 0.000; $X_6 - X_3$: PC = 0.458 i p = 0.000; $X_7 - X_3$: PC = 0.421 i p = 0.000.

Considering that there is a considerable number of variables with acceptable level of correlation and statistical significance ($p \leq 0.05$), it was concluded that the MLRA approach should be considered as the adequate tool for modelling of investigated process. For the purpose of MLRA analysis, the assembly of 500 input and output data sets was divided into two groups. First group consisted 350 (70%) of randomly selected data lines, and it was used for training of the model, while the second group consisted 150 (30%) remaining data lines from the starting data base and it was used for testing of the model.

Linear dependence of degree of analysed solution decomposition (Y) on influencing parameters of the technological process (X_1-X_7) was obtained using SPSS software application Version 17.0 (SPSS Inc, 2010). The complete linear model, developed during training of the model, is as follows:

$$Y = 102.864 - 0.044 \cdot X_1 - 23.108 \cdot X_2 + 1.817 \cdot X_3 + 0.140 \cdot X_4 - 0.297 \cdot X_5 - 0.136 \cdot X_6 + 0.027 \cdot X_7 \ (R^2 = 0.670) \quad (1)$$

The results of the ANOVA tests of developed model are presented in Table 3. Significant F statistics (see Table 3) is indicating that using the model is better then guessing the mean. Also, the significance value of the F statistic is less than 0.05 which means that the variations explained by the model are not due the chance. Regression displays information about the variation accounted for by the model, residual displays information about the variation that is not accounted for by the model. The ratio of regression to residual is 67%: 33%, advocating that 67% of the dependent variable (Y) values are explained by the model.

Table 3. Results of ANOVA^{a,b} test performed during training of the model

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	1772.337	7	253.191	99.341	0.000[a]
Residual	874.209	342	2.549		
Total	2646.546	349			

a. Predictors: (Constant), X7, X4, X6, X2, X1, X5, X3

b. Dependent Variable: Y

Results describing MLRA model summary, for the training phase, are presented in the table 4.

The multiple correlation coefficient (R) is presenting the linear correlation between the observed and model-predicted values of the dependent variable. Its large value (0.818) indicates a strong relationship. R square, the coefficient of determination, is the squared value of the multiple correlation coefficient.

It shows that about 67% of variation in Y is explained by the model, as already indicated by the regression to residual ratio.

As the future measure of the strength of the model fit, we compared the standard error of the estimate in the model summary table (Table 4) to the standard deviation of Y reported in the descriptive statistics table (Table 1). Without prior knowledge of the X_1– X_7 values, our best guess for the Y would be about 46.65% with standard deviation of 2.77. With the MLRA model the error of our estimate is considerably lower, about 1.59%. Considering above model parameters, the model developed according to MLRA seemed highly acceptable for prediction of the gibbsite crystallisation under the industrial conditions.

On the other hand, after running the collinearity analysis of the models coefficients, the results obtained showed that there might be a problem with multicollinearity of the model. For most predictors, the values of the partial and part correlations drop sharply from the zero-order correlation (see Table 5). This means, for example, that much of the variance in Y that is explained by X_2 is also explained by other variables. The tolerance is the percentage of the variance in a given predictor that cannot be explained by the other predictors. Thus, relatively small tolerances in case of predictors X_3, X_5 and X_7 show that more than half of the variance in a given predictors can be explained by the other predictors. A variance inflation factor (VIF) greater than 2 is usually considered problematic and in the example modelled in this paper; this is the case for predictors X_3, X_5 and X_7. Also, important factor of the collinearity analysis is condition index. Values of condition index greater then 15 indicate a possible problem, with collinearity greater than 30 a serious problem. Five of these indices were larger then 30, in case of this example, suggesting a very serious problem with collinearity. We tried to fix this collinearity problem by running the regression using stepwise method of model selection, hoping that most of the predictors will remain in the final model. Unfortunately, this did not improve the collinearity situation.

Table 4. MLRA summary^{a,b} of the model developed during the training phase

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.818^{a}	0.670	0.663	1.5965

a. Predictors: (Constant), X7, X4, X6, X2, X1, X5, X3

b. Dependent Variable: Y

However, after developing the model in the training stage, validation of the model was performed in the testing stage using the second part of the data base (total 150 vectors). During the testing phase of the MLRA model, calculated coefficient of determination (R^2) was slightly increased in comparison to the testing phase and now it equals: 0.731. Figure 1 illustrates comparative presentation of the measured and the values calculated using the MLRA

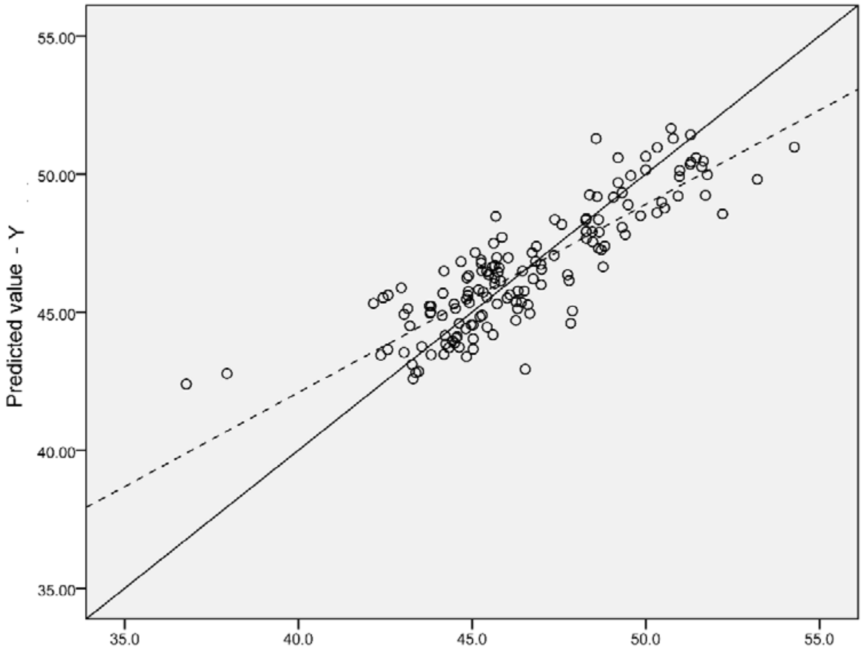


Fig. 1. Dependence between calculated and measured values of the caustic sodium aluminate solutions degree of decomposition (- ideal position; --- regression lines; o - valuescalculated using MLRA model in the testing stage)

approach for the investigated process. As presented in previous text, better fit was obtained on the test set than on the training set. This suggests that most of the extreme points that are more difficult to model are in the training set. The selection of the variables for the training and the testing stage was performed using random number generator and it was not subjectively influenced. Also, all the data lines were examined for potential outliers before the MLRA. Strong extreme behaviour of the variables was not detected.

Table 5. Results of collinearity analysis of the MLRA model

Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	X_1	-0.158	-0.043	-0.025	0.834	1.200
	X_2	0.072	-0.375	-0.233	0.702	1.425
	X_3	0.416	0.433	0.276	0.415	2.407
	X_4	-0.078	0.098	0.057	0.894	1.119
	X_5	-0.699	-0.452	-0.292	0.434	2.303
	X_6	-0.266	-0.484	-0.317	0.579	1.726
	X_7	0.649	0.161	0.094	0.443	2.256

This way, obtained results are the true represent of the investigated process.

4.1. Artificial neural networks

The ANN used in the model development is depicted in Fig.2. As shown, this network consists of three layers of nodes. The layers described as input, hidden and output layers, can generally comprise (i), (j) and (k) number of processing nodes, respectively. Each node in the input (hidden) layer is linked to all the nodes in the hidden (output) layer using weighted connections. In addition to the (i) and (j) number of input and hidden nodes, the ANN architecture also houses a bias node (with fixed output + 1) in its input and hidden layers and they provide additional adjustable parameters (weights) for the model fitting. The number of the nodes (i) in the ANN network input layer is equal to the number of inputs in the processwhereas the number of output nodes (k) equals the number of the process outputs. However, the number of hidden nodes (j) is an adjustable parameter magnitude of which is determined by issues, such as the desired approximation and generalization capabilities of the network model (Zeng, et al., 1997).

The back propagation algorithm modifies network weights to minimize the mean squared error between the desired and the actual outputs of the network. Back propagation uses supervised learning in which the input, as well as desired outputs are controlled and selected (Eberhart and Dobbins, 2002).

The use of ANN usually comprises threephase. First is the training phase which is facilitated on 70 to 80% randomly selected data from the starting data set. During this phase the correction of the weighted parameters of the connections is achieved through necessary number of iterations, until the mean squared error between the calculated and measured outputs of the network is minimal. During the second phase, the remaining 20 – 30% of the data is used for testing of the ”trained” network. In this phase, the network is using the weighted parameters determined during the first phase. This 20 – 30% of the data, excluded during the learning of the network, is now incorporated in it as a new input values X_i which is then transformed to the new outputs Y_i. The third phase is the validation of the network on completely new data set. This data set is usually consisting of the data from new experimental measurements of the same process. The validation phase is presenting the final level of successful or unsuccessful predicting using the network developed in the previous two stages, on future database (Živković et al., 2009; Liu et al., 2009).

Same as in the MLRA procedure, the assembly of 500 input and output data sets was divided into two groups. First group consisted 350 (70%) of randomly selected data lines, and it was used for training of the network, while the second group consisted 150 (30%) remaining data lines from the starting data base and it was used for testing of the network.

For development of relational ANN configuration we used previously defined input parameters $X_1 – X_7$ and output parameter Y (degree of decomposition of sodium aluminate solution), as the elements of the network architecture, Figure 2.

The appropriate number of neurons in the hidden layer was determined by training several networks. This is necessary because of the fact that too low number of neurons in the hidden layer produces high training and

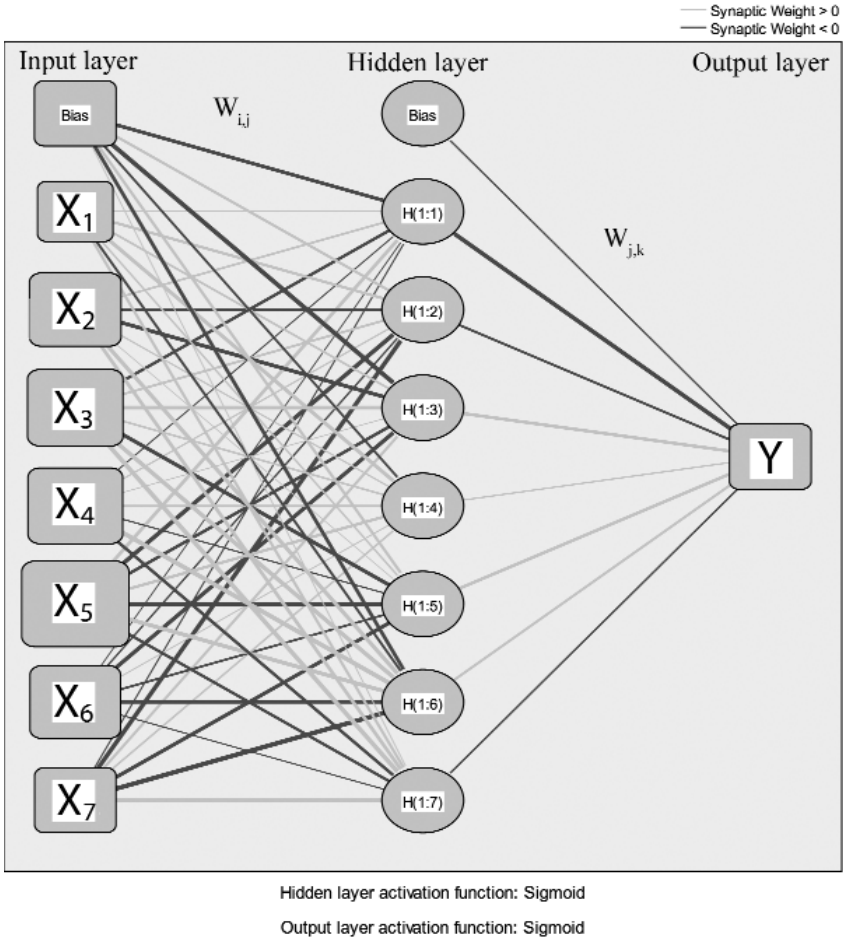


Fig. 2. The ANN architecture for determination of the degree of decomposition of the sodium aluminate solution as the function of input process parameters

testing errors due to under-fitting and statistical bias. On the contrary, too many hidden layer neurons lead to low training error, but high testing error, due to overfitting and high variance. Because of this, in this study, we used the iterative approach to determine the optimal number of hidden layer neurons, yielding minimum model prediction error on the "test data set". This way, we have tested 13 networks, ranging from 2 to 14 neurons in the hidden layer. The best results were obtained with the network architecture presented in Fig.2.

In the phase of training of the network, for each of the network architectures, necessary number of iteration was performed, until the error between the measured output of the decomposition process of industrial sodium aluminate solution Y- and calculated values was minimized and remained constant.

After developing of this kind of ”trained” network, testing stage was performed using the second part the data base (total 150 vectors). In this phase also, all 13 hidden layer structures were involved, until obtaining minimum model prediction error. The ANN structure presented in Fig. 2, with seven neurons in the hidden layer, resulted with minimum model prediction error.

For such network, obtained coefficient of

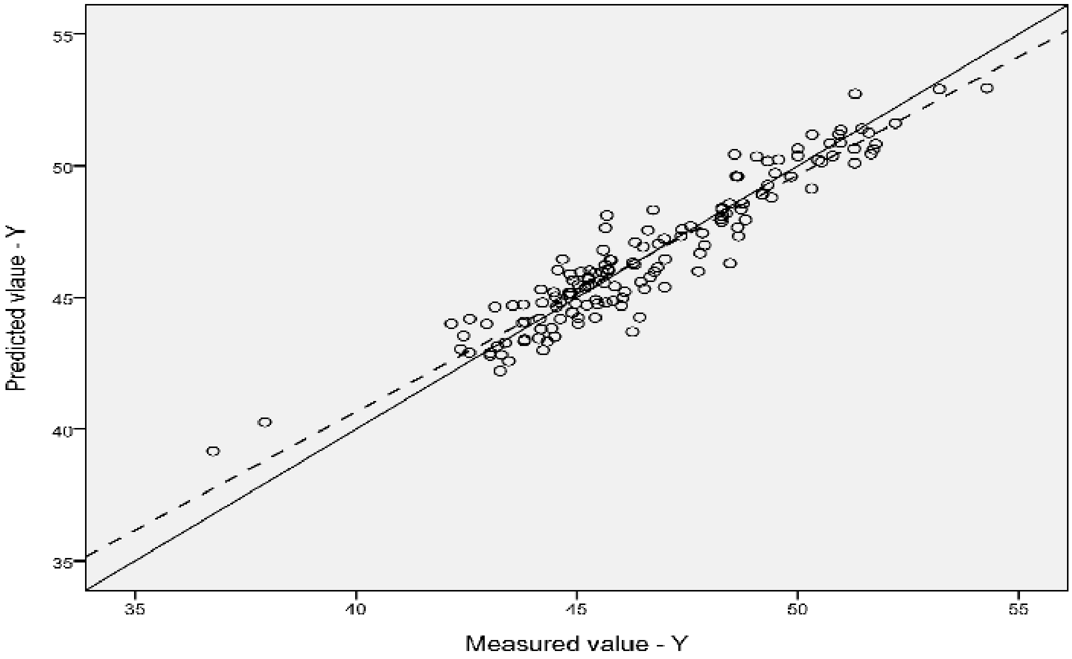


Fig. 3. Comparison of the measured and the values calculated using ANN for prediction of the degree of industrial sodium aluminate solution dissociation -Y

determination is $R^2 = 0.762$ for the training phase. During the ANN testing phase, calculated coefficient of determination (R^2) was slightly increased in comparison of the testing phase and now it equals: 0. 895. Figure 3 illustrates comparative presentation of the measured and the values calculated using the ANN approach for investigated process. The same situation happened as in the MLRA approach, meaning that better fit was obtained on the test set than on the training set. The explanation for this is the same as in the case of MLRA modelling, suggesting that most of the extreme points that are more difficult to model are in the training set.

5. CONCLUSIONS

Results obtained indicate that the industrial data collected in this study can be used for the purpose of predicting gibbsite crystallization. However, the model developed cannot be used for optimization of the process, since optimization requires cause and effect relationships of the data which can only be obtained through orthogonal design of experiments (DOE) and not from normal production data, because of the collinearity of the variables (see Table 2 and Table 5).

References

Dragićević, S, Bojić, M., (2009) Application of linear programming in energy management, Serbian Journal of Management 4 (2): 227 - 238.

Eberhart, R.C., Dobbins, R.W. (2002) Neural Network PC Tools: A Practical

Guide, Academic, New York.

Giraldo-Zuniga, A.D., Arevalo-Pinedo, A., Rodrigues, R.M., Lima, C.S., Feitosa, A.C. (2006) Kinetic drying experimental data and mathematical model for jackfruit slices, Cienc. Technol. Aliment., 5(2): 89-92.

Liu, D., Yuan, Z., Liao, S. (2009) Artificial neural network vs. nonlinear regression for gold content estimation in pyrometallurgy, Expert Systems with Applications, 36: 10397 – 10400.

Moroney, R.N. (1998) Spurious of virtual correlation errors commonly encountered in reduction of scientific data, Journal of Wind Engineering and Industrial Aerodynamics, 77&78: 543-553.

SPSS inc. PASW Statistics 18, Predictive Analysis Software Portfolio, www.spss.com

Taylor, C.F., Paton,N.W., Garwood, K.L. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data, Nature Biotechnology 21 : 247 - 254.

Weir G. (1991) Differential Equations - A Modeling Approach, Addison-Wesley, Hardcover.

Brown, C. (2007) Differential Equations - A Modeling Approach, Sage Publications.

Zeng, J., Yin, Z., Chen, Q. (1997) Intensification of precipitation of gibbsite from seeded caustic sodium aluminate liquor by seed activation and addition of crown ether, Hydrometallurgy, 89:107 – 116.

Živković, Ž., Mihajlović, I., Nikolić, Dj. (2009) Artificial neural network method applied of the nonlinear multivariante problems, Serbian Journal of Management, 4(2): 137 – 149.

Živković, Ž., Mitevska, N., Mihajlović, I., Nikolic, Đ. (2009) The influence of the silicate slag composition on copper loses during smelting of the sulfide concentrates, Journal of Mining and Metallurgy 45 B: 23-35.