PREDICTION OF PHOSPHORUS CONCENTRATION IN PRIMARY STEELMAKING

Luan Carlos de Oliveira

Cassius Zanetti Resende

Daniel Cruz Cavalieri

luandeol@gmail.com

cassius@ifes.edu.br

daniel.cavalieri@ifes.edu.br

Instituto Federal do Espírito Santo - Campus Serra

Rodovia ES 010, Km 6,5, 29173-087 - Manguinhos, Serra, ES, Brasil.

Leandro Rodrigues Ramos

Roberto Dalmaso

Henrique Silva Furtado

leandro.ramos@arcelormittal.com.br

roberto.dalmaso@arcelormittal.com.br

henrique.furtado@arcelormittal.com.br

Arcelor Mittal Tubarão

Av. Brigadeiro Eduardo Gomes, 930, Jardim Limoeiro, 29163970 - Serra, ES - Brasil.

Abstract. *The objective of this work was to predict the concentration of phosphorus of steel produced in Basic Oxygen Furnace (BOF), using real data composed of raw materials of the steelmaking process, the steel temperature, and result of the chemical analysis of the steel produced. Three techniques of data regression were applied, being PLS (Partial Least Square), PCR (Regression of Principal Components) and SVR (Support Vector Regression). The validation technique used was the k-fold cross-validation, and the performance indicator adopted was the RMSE (Root Mean Square Error). Another measure of efficiency of the model was based on the maximum allowed phosphorus value of the steel produced. Among the three techniques, a SVR with a Gaussian kernel obtained the best indicators, showing a good solution to the proposed problem.*

Keywords: *Regression, Machine Learning, SVR, PLS, PCR.*

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

1 INTRODUCTION

As globalization of the economy continues, steelmakers are faced with increasing competition. Product quality is often a differentiating factor among producers, driving the demand for the development of advanced tools to improve quality during the steelmaking process. Many machine learning algorithms have been used to improvethe productivity and the quality of such process, one can cite the works of (Wang et al., 2012), (Han; Liu, 2014), (Liu; Wang; Xiong, 2014) and (Han; Cao, 2015).

In certain applications, an elevated phosphorous concentration of the steel can be severely detrimental to the final product quality. For this reason, one of the goals of the steel refining process is to achieve a concentration of phosphorous less than the maximum allowed concentration. In this context, this work proposes the development of predictors for the concentration of phosphorous at the end of the BOF process, using the following supervised learning techniques: Partial Least Squares (PLS), Principal Component Regression (PCR), Support Vector Regression (SVR), and Artificial Neural Network (ANN). Since steel refinement is an aggressive process conducted at high temperature (>1600°C), there are limitations in measuring the variables involved, thus only 13 static process variables were used to train the algorithms. Although developed predictors use a reduced amount of input information, their accuracy rate was higher than the cited works.

The steelmaking process is divided into two main stages, the primary and the secondary refining. In a simple way, the primary refining is a method in which carbon-rich molten pig iron is made into steel. Blowing oxygen through molten pig iron lowers the carbon content of the alloy and changes it into steel. In the most common form of steel production, this process is carried out in equipment called an oxygen converter, or BOF (Basic Oxygen Furnace). Initially the converter is lined and receives scrap metal. Then it receives the liquid pig iron and with the metallic charge it receives the oxygen blow and inert gases for the partial oxidation of impurities and the reduction of carbon levels. At the end of this step a sample is collected to analyze the chemical composition and subsequent adjustments in the chemical composition, if necessary.

In secondary refining, the steel undergoes new chemical adjustments that will result in the desired final chemical composition. However, it should be emphasized that the final concentration of phosphorous is not be changed during in the secondary refining. This fact emphasizes the importance of the prediction of this specific chemical.

Returning to the primary refining, an important moment is when an intermediate sublance, with a sensor for measuring the temperature as well as the P-content of the bath, is dipped into the bath during the last third of the blowing process (Heinz, 2009). In this moment, samples of the steel in production are collected in order to be analyzed in laboratory to verify their chemical composition. The same process is repeated at the finish of the primary refining.

The steel sample, although necessary in the conventional model, is a limiting factor in productivity, whereas the results of the analysis are not obtained instantaneously and generate a delay in the end of the primary refining, especially when there is a need for readjustments in the chemical composition for suitability to the desired compositions. Thus, the objective of

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

the work is the prediction of phosphorus concentration in the batch in order to eliminate the need for intermediate chemical analysis in the steelmaking.

The study is organized in three sections. Section 2 discusses the adopted methodology, describing the procedures necessary for the development of the models and presents them in a succinct way the techniques used. Section 3 highlights the results achieved, as well as comments that aid in understanding. Finally, Section 4 contains the conclusions and some suggestions for future work.

2 MATERIALS AND METHODS

This section describes the methodology adopted in the development of the work to predict Phosphorus concentrations in steel produced.

2.1 Regression techniques

Three regression techniques were used to predict Phosphorus concentrations in steel: PLS (Partial Least Squares), PCR (Principal Component Regression) and SVR (Support Vector Regression). In relation to the PLS technique, two variants were used: SIMPLS and NIPALS, accounting for four models.

The above techniques are based on supervised learning, i.e., at the training stage, there is an associated variable (or dependent) for each set of explanatory (or independent) variables.

The descriptions of all of them are shown in the following subsections.

2.1.1 Partial Least Squares

The Partial Least Squares (PLS) method is based on latent variables, each of which is obtained by the linear combination of the independent variables (matrix X) or the dependent variable (matrix Y).

The objective of PLS is to construct a linear model of the form $Y = X\beta + \varepsilon$, where Y is a matrix containing the answers (dependent variables), X is matrix containing the independent or explanatory variables, β is the matrix containing the regression coefficients of the model and ε is the matrix of the residues.

The model based on partial least squares results in a score matrix, denoted by *T*. By maximizing the covariance between *X*, *T* and *Y*, one can capture the variance of the explanatory variables, and at the same time obtain correlations with the dependent variables (Schumann; Nolte; Zheng, 2013).

The matrix of scores T is formed by linear combinations between the original variables X and a matrix of weights W. With this, the original matrices can be decomposed as follows

$$X = TP' + F, \text{and} \quad (1)$$
$$Y = UC' + G. \quad (2)$$

In which T and U are, respectively, the scores matrices of X and Y, P' and C' are weight matrices and F and G are matrices of residuals of X and$Y$.

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

Since the scores of *X* are good estimators of *Y*, Eq. 2 becomes

$$Y = TP' + G. \tag{3}$$

The matrix *C'* is obtained by least squares by

$$C' = (T'T)^{-1}T'Y. \tag{4}$$

The combination of Eq. 3 and Eq. 4 results in

$$Y = TC' + E = XWC' + E, \tag{5}$$

where *E* is the matrix of residuals of estimation.

2.1.2 Principal Component Regression

The PCR method has some advantages over the partial least squares method, one of which is the ability to bypass problems related to the inversion of the weight matrix, shown in Eq. 4 and the collinearity between the explanatory variables.

The PCR employs the steps similar to PCA (Principal Component Analysis) and decomposes only the matrix *X*, resulting in the matrix of scores. With the results of the decomposition of the matrix *X*, the regression is performed, i.e., the PCR performs the regression of the matrix of scores and not of the original matrix *X*, as is done in PLS, and this way it is guaranteed that the matrix of scores is invertible.

From the point of view of collinearity, when the main components are decomposed, each one is an independent and additive component for representing the variability of the data.

The major disadvantage, however, is that the PCR application may result in loss of information to improperly chosen of the major components in modeling (Schumann; Nolte; Zheng, 2013).

2.1.3 Support Vector Regression

The SVR is a nonparametric regression learning technique, i.e., it is not necessary to know the probability distribution of the universe, or process, from which the data sample under analysis was obtained.

The objective of the SVR is to adjust the training matrices *X* and *Y* to a linear function in the format

$$f(X) = WX + B, \tag{6}$$

where *W* is the weight matrix and *B* is the error matrix.

The adjustment is made so that the point-to-point error is not greater than a given value ε. In this way, it is finding the weights that satisfy the condition presented in Eq.7:

$$|y_i - f(x_i)| < \varepsilon, \tag{7}$$

for all $i = 1, 2, \ldots, m$.

Therefore, the great advantage of SVR in relation to the other techniques mentioned is the fact that its training is equivalent to solving a quadratic optimization problem in which the solution is global and there is no local minimum (Laha; Ren; Suganthan, 2015).

2.2 Contents of the database

The database used in the modeling is composed of real data obtained from a steelmaking process. The independent variables are composed basically of inputs inherent to the primary refining of the steel in BOF, while the dependent variable is the final concentration of the Phosphorus. Hereafter, each set of independent variables associated with the respective dependent variable will be demonized by the term batch.

Given the aggressiveness of the steelmaking process, especially due to the high temperatures reached, there are limitations regarding its instrumentation and monitoring. Thus, only 13 independent variables were used in the modeling. Such independent variables are described in Table 1.

Table 1. Input variables used in modeling

Input	Description
IN_1	Temperature of the steel at the end of blow
IN_2	Temperature of the steel in the sublance
IN_3	Estimated carbon content in sublance
IN_4	Total blow Oxygen
IN_5	Oxygen measured in the sub lance
IN_6	Net pig iron rate
IN_7	Solid pig iron rate
IN_8	End-of-blow slag weight
IN_9	Iron ore added
IN_10	Dolomite added
IN_11	Briquette added
IN_12	Concentration of phosphorus in pig iron
IN_13	Concentration of Manganese in pig iron

The database contains 2049 baths and for each one of them it is still present the maximum concentration of Phosphorus allowed. This information was used, together with the final phosphorus concentration, to analyze the performance of the models.

Although the steel is composed of several different elements, in this work, the efforts were directed in the estimation of the final Phosphorus, since this element is directly associated to the hardness properties of the steel.

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

2.3 Collinearity analysis

In order to evaluate the relevance of the variables, the Pearson correlation coefficients were been measured. They were calculated between the independent variables X_i taken in pairs and between the independent variables X_i and the dependent variable Y.

Because they are extensive, the correlation values between the independent variables taken in pairs will not be shown. The correlations between independent variables X_i and the dependent variable Y are presented in Table 2. The correlation coefficient measures how much a variable can be estimated or explained from another, assuming values in the range between -1 and +1. The extreme values of the range indicate perfect colinearity, with -1 for perfect inversely proportional correlation and +1 for directly proportional perfect correlation. Finally, the null coefficient indicates statistical independence between the variables.

Ideally, for the good performance of the models, the correlation between the entries should be zero, that is, the absence of multicollinearity between the explanatory variables. This situation indicates that they exert individual and additive influences on the dependent variable (Ziegel; Eric R, 1999).

Table 2. Correlation coefficients

Input	Correlation with output
IN_1	0.4419
IN_2	0.3284
IN_3	0.0359
IN_4	-0.1377
IN_5	-0.0905
IN_6	0.0028
IN_7	0.0316
IN_8	-0.3592
IN_9	0.0612
IN_10	0.0851
IN_11	-0.0795
IN_12	0.0925
IN_13	-0.0156

2.4 Data conditioning

An essential step in modeling is the preprocessing of the base used in the training. In general, the purpose of conditioning is to eliminate or minimize the influence of situations such as missing inputs, outliers and to avoid redundancy of information. This conditioning reduces its computational complexity and makes parameter adjustment more flexible (García; Luengo; Herrera, 2016).

The first filtration performed on the base was the treatment of outliers. This procedure aimed first at the elimination of null values in the first eight entries. Due to the behavior of

the process, the independent variables IN_1 to IN_8 cannot assume null values and, consequently, samples with this characteristic were disregarded in the modeling. The presence of null values in these variables is due to the failures of sensors and instrumentation equipment, since the database represents measurements obtained in an industrial plant.

After this initial filtering, the 1.5×IQR (Interquartile Range) method was used as the data safety range (Riaz, 2015). In this way, the values of the independent variables outside this range were considered outliers and the corresponding batch were disregarded in the modeling. Although, eventually, values outside the range of 1.5×IQR were authentic information of the process, this methodology was chosen in order to establish a compromise relationship between the generalization capacity of the models and their accuracy.

Finally, the previously filtered batches underwent a process of normalization of the independent variables, since there were large discrepancies between the ranges of values. Normalization has as main objective to assign the same degrees of importance between inputs and not to carry out the weighting of the model based on only some of them, i.e., to avoid problems of bias or polarization of the models.

3 EXPERIMENTAL RESULTS

The results presented in this section refer to 204 batches submitted to the regression models. The test partition used corresponds to 10% of the original base, without any kind of filter or sample removal being applied. That is, the test partition consists of authentic samples of the process and covers the various situations of lack of data, presence of noise and outliers. This characteristic is important to evaluate the robustness of the models.

In addition to the methodologies used in Section 2, for the purposes of practical validation, the models had their performance confronted with an ANN (Artificial Neural Network) model that is already operating in the industrial plant.

In the calibration and validation of the models, k-fold was used to ensure good generalization capacity and to avoid overtraining (Liang; Von Davier, 2014). Thus, from the original database ten partitions were created, each one containing randomly samples, which means that the partitions used in the validation are mutually exclusive, i.e., they do not contain any batch in common.

The measure of efficiency of the models was based on the RMSE (Root Mean Square Error) and on the maximum phosphorus concentration allowed by each batch. In order to define the types of errors and successes, one defines

- True positive (TP): The predictions are smaller than the maximum permissible concentration and whose laboratory analysis of the steel proved to be true prediction of the model;

- True negative (TN): The predictions are larger than the maximum allowed concentration and whose laboratory analysis of the steel proved to be true prediction of the model;

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

- False positive (FP): The predictions that the model estimates to be smaller than the maximum allowed concentration, but whose laboratory analysis of the steel revealed the opposite. This is the most serious case for the prediction error;

- False negative (FN): The predictions that the model estimates are larger than the maximum allowed concentration, but whose laboratory analysis of the steel revealed otherwise.

The sum of the TP and the TN is said to be correct. This indicator is used in decision-making for direct tapping, i.e., to proceed with the batch in the refining process or to readjust the inputs for corrections of the chemical composition.

The evaluation of the residuals (ε) was used as additional validation of the models. This value corresponds to the difference between the estimated value and the real value, being important to show the non-bias in the modeling. In short, the residues should result in a random distribution, so the model is said to be non-biased. For this purpose, the Residual Normality Test was performed (Antonakis; Dietz, 2011).

3.1 RMSE

Table 3 shows the RMSE values for each model. It should be noted that although the regression models have been slightly better than the ANN, there are proximity in their mean error values, around 0.0023.

Table 3 – Averages of RMSE values for the 10 partitions

Model	RMSE
ANN	0.0033
NIPALS	0.0024
PCR	0.0023
SIMPLS	0.0024
SVR	0.0023

Table 4 shows the results adopting a tolerance interval of ± 20 ppm around the correct values of the concentration of the phosphorus, i.e., considering correct the estimates in which the error modules resulted in values lower than 20ppm. As a result, the SVR resulted in a hit rate of 65%.

Table 4 – Efficiency of the models in relation to the real Phosphorus + tolerance

Model	Efficiency (%)
ANN	58.35
NIPALS	62.25
PCR	63.57
SIMPLS	63.57
SVR	65.81

The Fig. 1 shows the adjustment of the correct Phosphorus estimates around the actual concentrations obtained by laboratory analysis.

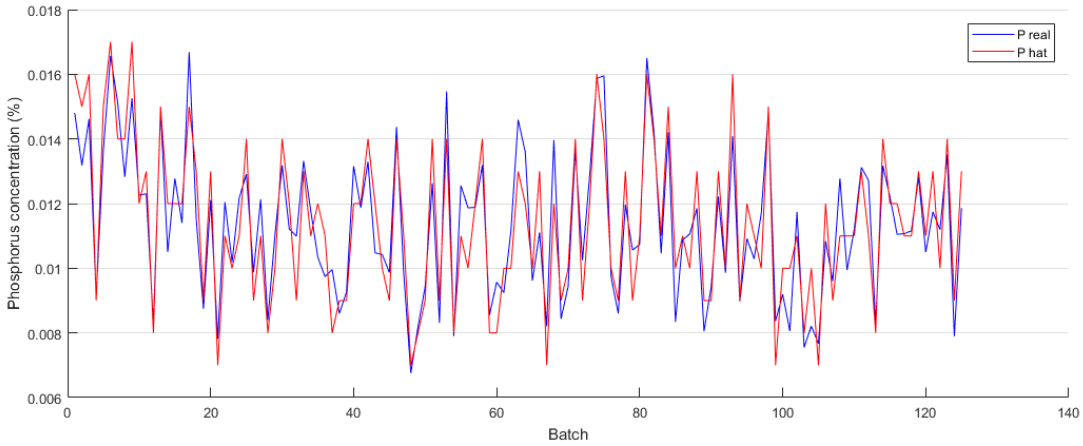

Figure 1. Correct Phosphorus estimates in data test

3.2 Maximum phosphorus concentration allowed

Table 5 presents the estimates from the point of view of the maximum permitted phosphorus concentration in each batch. Again, among the techniques, the SVR presented the best performance, around 91% hit.

Table 5 – Models efficiency in relation to the maximum phosphorus

Model	TP (%)	TN (%)	FP (%)	FN (%)	TP+TN (%)	FP+FN (%)
ANN	84.51	3.64	6.03	5.81	88.15	11.84
NIPALS	86.30	5.00	4.78	3.91	91.30	8.69
PCR	87.01	4.61	5.16	3.20	91.62	8.36
SIMPLS	87.01	4.61	5.16	3.20	91.62	8.36
SVR	87.33	4.40	5.38	2.88	91.73	8.26

It can be seen that the models present very close results, but with a slight advantage of SVR, evaluated both in relation to the real concentration and in relation to the maximum allowed rate. It is also noted that the results of PCR and SIMPLS were the same. This is justified by the fact that both have similar procedures in the execution of their algorithms, being one of the steps in the dimensionality reduction of the samples under analysis.

In the graph of Fig. 2 is shown the dispersion of correctly estimated samples by the SVR model, from the point of view of the actual concentration of Phosphorus.

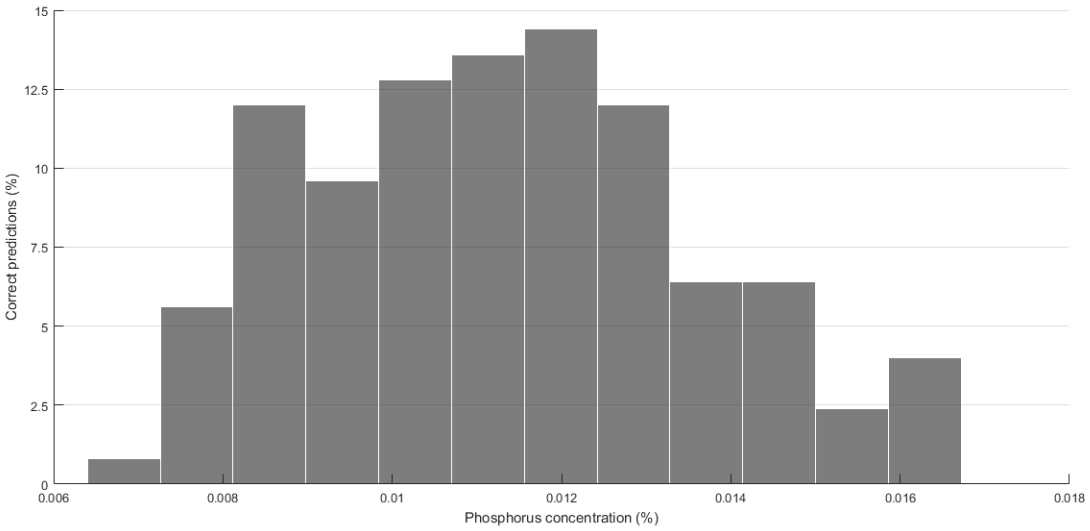

Figure 2. Corrects SVR estimates

3.3 The evaluation of the residuals (ε)

In order to ensure the non-bias of the estimates, the residuals of the estimates were evaluated. The graphs in Figs. 3 and 4 show, respectively, the residues obtained in the test partition and their probability distribution. One can note that they are scattered randomly, which, graphically, shows signs of non-bias. Moreover, they are centered on zero, with a mean of $-1.9 * 10^{-4}$ and normally distributed.

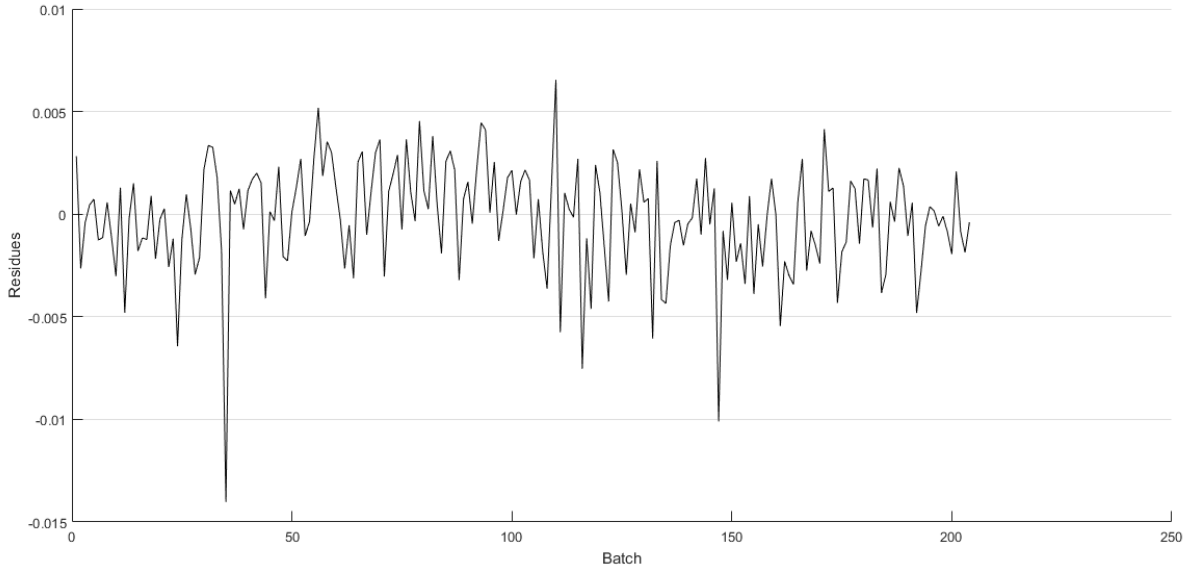

Figure 3. SVR residues

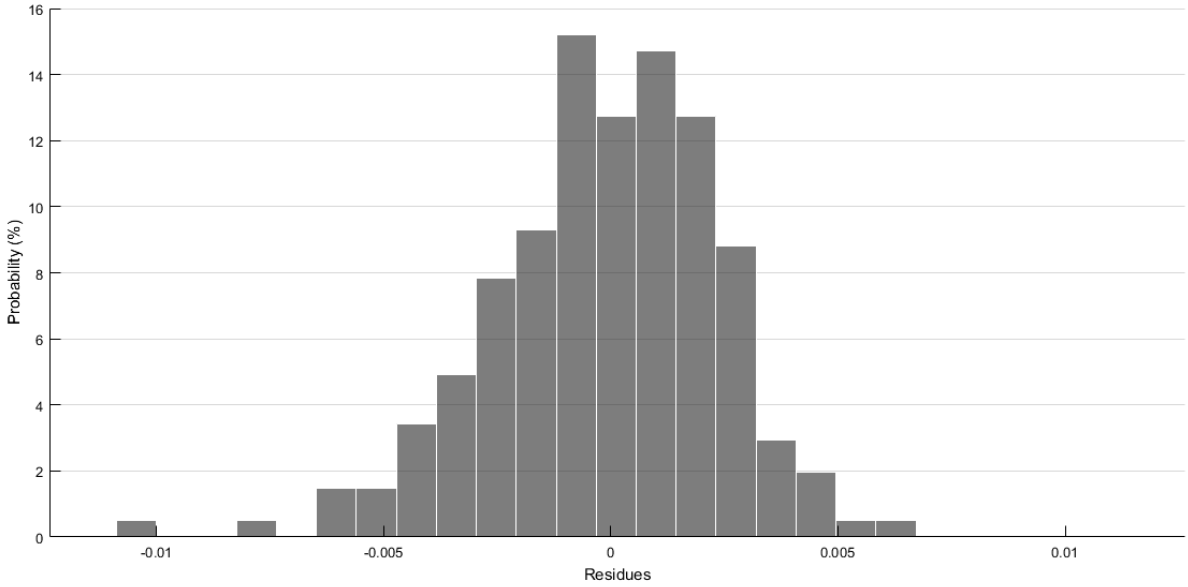

Figure 4. Residues probability distribution

Finally, Table 6 presents the Pearson correlation coefficients between the residuals and the independent variables. The presence of correlation, although weak, evidences statistical dependence between the factors not inserted in the model and the independent variables. This means that explanatory variables not included in the modeling do not present statistical independence with the 13 inputs used and do not have purely independent and additive contributions on the dependent variable (final concentration of Phosphorus).

This contradicts the premise of statistical independence between the independent variables and explains, in addition to the few variables used in the modeling, the performance of the models for the estimates in relation to the actual concentrations of Phosphorus.

Table 6. Correlation coefficients between inputs and residues

Input	Correlation coeficients
IN_1	-0.0659
IN_2	-0.1369
IN_3	0.0349
IN_4	0.0699
IN_5	-0.0052
IN_6	-0.0618
IN_7	0.0766
IN_8	0.1670
IN_9	-0.0684
IN_10	-0.0966

IN_11	0.1233
IN_12	-0.0522
IN_13	0.0095

4 CONCLUSION

The present work had as objective the development of models for the estimation of phosphorus concentration at the end of the primary refining of the steel in basic oxygen furnace.

Among the three regression techniques used, SVR presented the best performances, with a hit rate of around 91%, adopting as threshold the maximum permissible concentration of phosphorus for direct tapping.If compared with real concentrations plus the tolerance of ±20ppm, the estimates showed low performance, about 65% for SVR. However, given the analyzes performed, it is inferred that the hit rates, as well as the estimation residuals can be improved with the addition of more independent variables as inputs to the models, a suggestion for future work. Given the complexity of refining, the use of only thirteen independent variables is insufficient to have a good representation of the process. This was confirmed by the correlation coefficients, which indicated that among the thirteen independent variables, only three of them had a more significant degree of correlation with the dependent variable.

Finally, the insertion of dynamic refining data in the modeling, such as the oxygen blow profile, the relative height between the blower tip and the liquid metal surface, and the blow profile of the inert gases, can still be explored. It is believed that the addition of information of this nature can improve the accuracy of the models.

REFERENCES

Antonakis, & Dietz. (2011). Looking for validity or testing it? The perils of stepwise regression, extreme-scores analysis, heteroscedasticity, and measurement error. Personality and Individual Differences, 50(3), 409-415.

García, Luengo, & Herrera. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, *98*, 1-29.

Han, & Cao. (2015). An improved case-based reasoning method and its application in endpoint prediction of basic oxygen furnace. Neurocomputing, 149, 1245-1252.

Han, & Liu. (2014). Endpoint prediction model for basic oxygen furnace steel-making based on membrane algorithm evolving extreme learning machine. Applied Soft Computing Journal, 19, 430-437.

Han, & Zhao. (2011). Dynamic control model of BOF steelmaking process based on ANFIS and robust relevance vector machine. Expert Systems With Applications, 38(12), 14786-14798.

Heinz D. Unbehauen (2009). Control Systems, Robotics And Automation – Volume XIX: Industrial Applications of Control Systems-II, 43.

Laha, Ren, & Suganthan. (2015). Modeling of steelmaking process with effective machine learning techniques. Expert Systems With Applications, 42(10), 4687-4696.

Liang, T., & Von Davier, A. (2014). Cross-Validation. Applied Psychological Measurement, 38(4), 281-295.

Liu, Wang, & Xiong. (2014). Basic oxygen furnace steelmaking end-point prediction based on computer vision and general regression neural network. Optik - International Journal for Light and Electron Optics, 125(18), 5241-5248.

Riaz, M. (2015). On Enhanced Interquartile Range Charting for Process Dispersion. Quality and Reliability Engineering International, 31(3), 389-398.

Schumann, Nolte, & Zheng. (2013). Comparison of partial least squares regression and principal component regression for pelvic shape prediction. Journal of Biomechanics, 46(1), 197-199.

Shao, Yanming, Zhou, Muchun, Chen, Yanru, Zhao, Qi, & Zhao, Shuan. (2014). BOF endpoint prediction based on the flame radiation by hybrid SVC and SVR modeling. Optik: International Journal for Light and Electron Optics, 125(11), 2491-2496.

Wang, Xu, AI, & Tian. (2012). Prediction of Endpoint Phosphorus Content of Molten Steel in BOF Using Weighted K-Means and GMDH Neural Network. Journal of Iron and Steel Research International, 19(1), 11-16.

Ziegel, Eric R. (1999). Applied Multivariate Statistical Analysis Review. *Technometrics*, *41*(1), 81-82.