PRINCIPAL COMPONENT REGRESSION MODEL OPTIMIZATION OF STEEL IN AN ANNEALING HEATING FURNACE ON HOT-DIP GALVANIZING LINE

Badreddine Ifrah[1], Abdelkhalek Cheddadi[2]

[1]Maghreb Steel, Laboratory Thermal Systems and Real Flows, Mohammadia School of Engineers, University of Mohamed V, Rabat, Morocco, badreddine.ifrah@maghrebsteel.ma
[2]Laboratory Thermal Systems and Real Flows, Mohammadia School of Engineers, University of Mohamed V, Rabat, Morocco, cheddadi@emi.ac.ma

Abstract

Predicting the temperature of the steel strip in an annealing heating furnace on hot-dip galvanizing line (HDGL) is important to ensure the physical properties of the processed material. The objectives are the optimization of the quality by identifying the minority of the processes and the activities responsible for the majority of the energy costs and for the most effective factors in design of the process which support continuous and continual improvement is recently discussed from different points of view. In this study, we examined the quality engineering problems in which several characteristics and factors are to be analyzed through a simultaneous equations system. The solution is presented for modeling and optimizing in furnace of annealing the line of galvanization industry metal, using by principal components regression model.

Keywords: *Hot-dip galvanizing line, Data mining, Principal components regression, Modeling, optimization.*

1. Introduction

A continuous hot dip galvanizing line (CHDGL) is a well-known steel industrial process consisting of several stages. The initial product is a steel coil which is the result of prior rolling processes. First, the coil is unwound and runs through a series of vertical loops within a continuous annealing furnace (CAF). This thermal treatment is fundamental to improving the properties of the steel. Then, the steel strip continues through a molten-zinc coating bath, followed by an airstream wipe that controls the anti-corrosion coating thickness. Finally, the strip passes through a series of auxiliary processes which wind the product back into a coil or cut it into flat products [1].

One of the most important stages in a CHDGL is the annealing treatment of the steel strip prior to zinc immersion (Fig. 1). When a steel strip receives a non-uniform heat treatment, it may cause inadequate steel properties, inconsistency in the quality of the coating layer, and other additional problems, such as surface oxidation or carbon contamination [2].

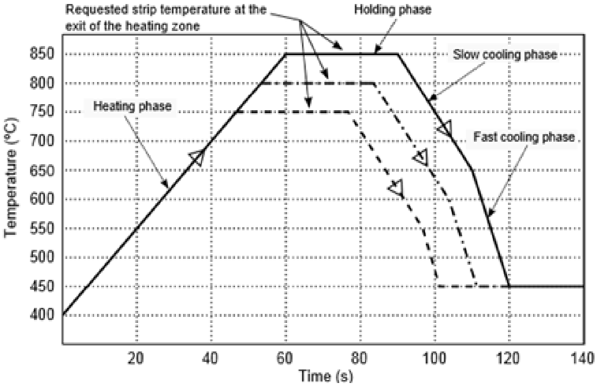

Fig 1: Example of three annealing treatment profiles. TMPHEAT is the strip temperature at the exit of the heating zone.

Nowadays, many operating systems for CHDGL are based on data driven models that predict the optimal CAF settings for each type of coil. These models usually establish furnace temperatures and strip velocity setting according to the dimensional characteristics of each coil, steel chemical composition, pre-established annealing curves of each type of steel, and the plant's operational requirements [3]. Thus, several studies report reliable models for predicting galvanizing set points [4].

Making decisions about complex problems involving process optimization and engineering design is strongly dependent on well identification of effective factors [5]. The most common goal of the factory owner is to achieve better quality in the final product by means of process improvements. The significance and relevance of optimizing the existing control models is even greater in the open-loop control systems or in those governed by computational methods dependent on adjustable parameters [6].

This paper reviews some typical industrial environments and focuses on some parts of them in order to show the real interest of these improvements. We will identify some difficulties in obtaining these improvements and find the interrelations between input parameters and output quality characteristics. According to the literature, many works have been conducted on using Principal Components Regression (PCR) approaches that may be viewed in terms of alternative formulations of the optimization problems. PCR are tools that are used to systematically examine different types of problems that arise within, e.g., research, development and production.

So, essentially regression on principal components is necessary, because we want to develop a stochastic model on the dependent variables using all these independent variables and the derived principal components as we derive from the first set, we do the principal component analysis on the independent variables and these principal components are used as independent variables in the regression [7].

It is reasonable to assume that the outcome of an experiment is dependent on the experimental conditions. This means that the result can be described as a function based on the experimental variables,

$$\underline{Y} = X\underline{B} + \underline{e}$$

Where $\underline{Y}$ is the dependent variable, X represents the independent variables, $\underline{B}$ is the regression coefficients to be estimated, and $\underline{e}$ represents the errors or residuals.

In any experimental procedure, several experimental variables or factors may influence the result. A screening experiment is performed in order to determine the experimental variables and interactions that have significant influence on the result, measured in one or several responses [8].

This method gives results but it is very expensive in time because it inevitably requires the realization of a great number of experiments. This is why it is important to help the scientist to achieve his experiments with principal components regression methods. PCR makes it possible to collect, summarize and present data so as to optimize the way to implement next experiments. By using experimental design, the scientist knows how to plan experiments. This experimental step will help him to structure his research in a different way, to validate his own assumptions, with better understanding the studied phenomena, and to solve the problems.

To help and answer this problem, a method of modeling by PCR of systems was implemented; the experiment cannot be anything, it has to supply the wished information. This experimental approach is going to help the experimenter to structure his search in a different way, to confront and to validate his own hypotheses, to understand better the studied phenomena and to solve the problems. The success of this methodology is partially bound to the needs for competitiveness of companies, but also to the desire to change the way of making experiments.

2. DATA MINING PROCESS

In this research project, registers as part of rationalization and mastery of atmospheric gases (nitrogen, hydrogen) and fluid (Liquefied Petroleum Gas) of the galvanizing line in Maghreb Steel. The Hot Dip Galvanizing line has changed from cold rolled material to a strip with dedicated technological properties. Several important process steps are necessary to transform the strip to the wanted state.

Data acquisition is obtained from the computer processing area based on the historical data continuously generated during the galvanizing process. The variables are selected according to their relevance to the furnace Heating Zone [2] [9].

The database consists of 20,250 records obtained from a galvanizing process involving 750 coils.

All variables are measured every 100m along the strip. The strip velocity is measured in the center of the furnace, and it is reasonable to assume that the strip maintains the same velocity throughout the Heating Zone. The relevant variables and their abbreviations can be found in table 1.

We have reviewed existing models and strategies for modelling technological parameters and coating appearance. It was also clear that control strategies over some parameters could provide us with information about the quality of the final product, in a continuous way, like information arising from skinpass. Special effort was dedicated to reviewing metallurgical works describing known laws between parameters at this level with both, mechanical properties from one side and operational conditions from the other side, and studies on the optimization of industrial processes employing PCR Techniques.

We will use the principal component analysis and then fit this regression equation. So, on the principal component for the principal component analysis the first thing we do is to standardize the original data that is a on the independent variables and also center the dependent variable data then we have to carry out the principal component analysis.

This and other examination stated the coherence between the applied methods of data analysis and the existing knowledge about relations between process and material parameters.

In order to answer the studied problem, we followed the steps below [10]:

The need of statistical robust methods to reinforce data cleaning processes carried out at the beginning of the data treatment was started in laboratory and plant studies.

Defined the major requirements for the availability and quality of the data used in the foreseen analysis. This includes the examination of all data available in the database concerning meaning.

Integrated pre-processing functions of PCR were used like outlier test and outlier elimination, grouping of data related to certain parameters and generation of training and validation samples for modelling.

Improvement of physical modelling of material structure evolution and adjustment of the technological parameters with help of continuous measurement.

Data-based modelling of technological parameters founded on operational process variables (length related data base and continuous measurement).

Data-based modelling of technological parameters founded on operational process variables (piece related database).
Data-based modelling of zinc coating appearance.

Integration of models into a framework for open-loop quality control strategies of hot dip galvanized sheet.
The following sketch will explain the interrelations of these different techniques and procedures applied in the project.

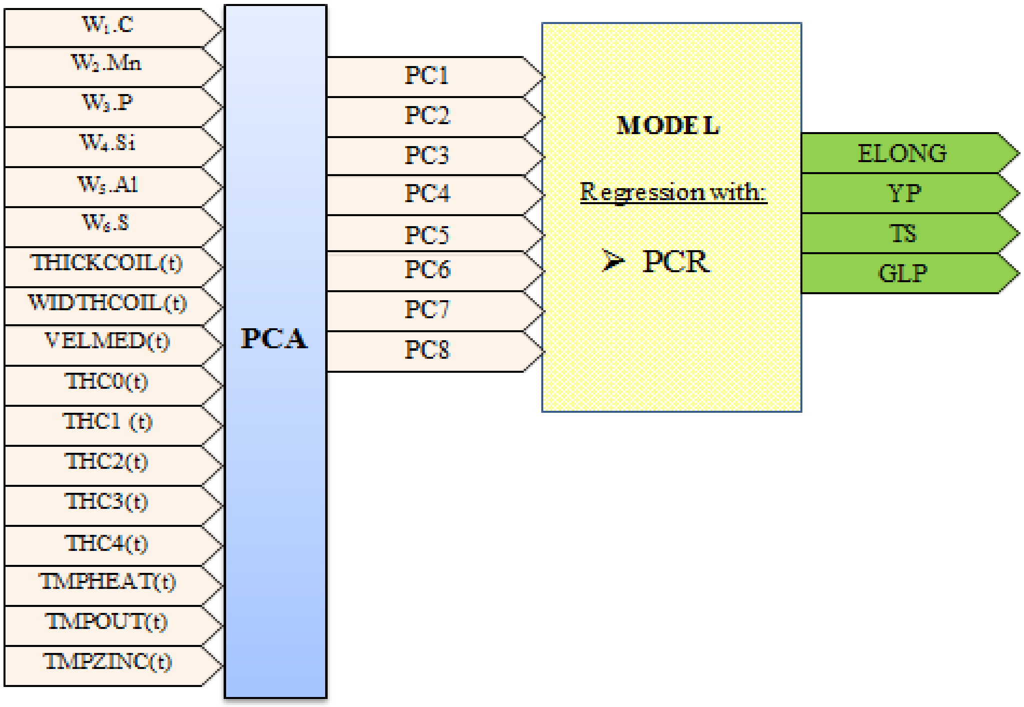

Fig 2: The methodology used in that application

3. The Proposed Method

Principal Components Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, principal components regression reduces the standard errors. It is hoped that the net effect will give more reliable estimates.

In a first step to test the models some simple known relationships between process variables and product quality defining quantities have to be found. It is possible to examine the relations of several variables at the same time by creating a graph Dispersal of type matrix. The created matrix presents all the pairs of variables of the list elaborated by the researcher. The obtained matrix contains so many rows and columns as they are variably declared in the list. Every cell of the matrix represents the graph of cloud of points of the pair of variables created by the intersection of the column and the row [11].

$$\underline{Y} = X\underline{B} + \underline{e}$$

Note that since the variables are standardized

$$X'X = R$$

Where R is the correlation matrix of independent variables. To perform principal components (PC) regression, we transform the independent variables to their principal components. Mathematically, we write:

$$X'X = PDP' = Z'Z$$

Where D is a diagonal matrix of the eigenvalues of $X'X$, P is the eigenvector matrix of $X'X$, and Z is a data matrix (similar in structure to X) made up of the principal components. P is orthogonal so that.

$$P'P = I$$

We have created new variables Z as weighted averages of the original variables X . This is nothing new to us since we are used to using transformations such as the logarithm and the square root on our data values prior to performing the regression calculations. Since these new variables are principal components, their correlations with each other are all zero. If we begin with variables $X1, X2$ and $X3$, we will end up with $Z1, Z2$ and $Z3$.

Severe multicollinearity will be detected as very small eigenvalues. To rid the data of the multicollinearity, we omit the components (the z's) associated with small eigenvalues. Usually, only one or two relatively small eigenvalues will be obtained. For instance, if only one small eigenvalue were detected on a problem with three independent variables, we would omit $Z3$ (the third principal component) [12].

When we regress Y on $Z1$ and $Z2$, multicollinearity is no longer a problem. We can then transform our results back to the X scale to obtain estimates of B. These estimates will be biased, but we hope that the size of this bias is more than compensated for by the decrease in variance. That is, we hope that the mean squared error of these estimates is less than that for least squares. Mathematically, the estimation formula becomes

$$\hat{\underline{A}} = (Z'Z)^{-1} Z'\underline{Y} = D^{-1}Z'\underline{Y}$$

Because of the special nature of principal components. Notice that this is ordinary least squares regression applied to a different set of independent variables. The two sets of regression coefficients, A and B, are related using the formulas:

$$\underline{A} = P'\underline{B}$$

and

$$\underline{B} = P\underline{A}$$

Omitting a principal component may be accomplished by setting the corresponding element of $\underline{A}$ equal to zero. Hence, the principal components regression may be outlined as follows:

1. Complete a principal components analysis of the X matrix and save the principal components in Z.
2. Fit the regression of Y on Z obtaining least squares estimates of A.
3. Set the last element of A equal to zero.
4. Transform back to the original coefficients using $\underline{B} = P\underline{A}$.

4. Results and discussion

In a first step to test the models some simple known relationships between process variables and product quality defining quantities have to be found. The following figure shows a correlation matrix and the result of a non-linear correlation analysis by the component plane of the Self-Organizing Map. The correlation matrix shows low and similar values between factors that are influential in the annealing. There is a significant correlation among two or more variables among the independent variables. So, this correlation needs to be addressed when we are developing a regression relationship. We standardize all the independent variables and generate the vector x.

Here are the results of reliability and Exploratory PCA for the eighteen variables. Note too that if overall the variables don't correlate, signifying that the variables are independent of one another (and so there aren't related clusters which will correlate with a hidden factor), then the correlation matrix would be approximately an identity matrix. We can test Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy, and off Bartlett's Test of Sphericity [13].

TABLE 1 RESULTS OF KAISER-MEYER-OLKIN (KMO) MEASURE OF SAMPLING ADEQUACY/BARTLETT'S TEST OF SPHERICITY

Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy.		0,727
Bartlett's Test of Sphericity	Approx.Chi-Square	1159,108
	df	136
	p-value	< 0,0001
	alpha	0,05

The Bartlett's test checks if the observed correlation matrix $R = \left(r_{ij}\right)_{(p\,x\,p)}$ diverges significantly from the identity matrix (theoretical matrix under H0: the variables are orthogonal). The PCA can perform a compression of the available information only if we reject the null hypothesis.

In order to measure the overall relation between the variables, we compute the determinant of the correlation matrix $|R|$. Under H0, |R| = 1; if the variables are highly correlated, we have $|R| \approx 0$.

The Bartlett's test statistic indicates to what extent we deviate from the reference situation $|R| = 1$. It uses the following formula.

$$\chi^2 = -(n-1-\frac{2p+5}{6})\times \ln|R|$$

Under H_0, it follows a χ^2 distribution with a $\left[p\times(p-1)/2\right]$ degree of freedom.

If two variables share a common factor with other variables, their partial correlation (a_{ij}) will be small, indicating the unique variance they share.

$$a_{ij} = \left(r_{ij} \bullet 1, 2, 3, \ldots k\right)$$

Interpretation:

The following table 2 gives information about two hypotheses of factor analysis. From the following table, we find out that sample sufficiency index KMO by Kaiser-Meyer-Olkin, which compares the sizes of the observed correlation coefficients to the sizes of the partial correlation coefficients for the sum of analysis variables is 72,7%. In addition, the control of sphericity (Bartlett's sign<0.001) proved that the principal component analysis has a sense. Through this analysis, supposition test of sphericity by the Bartlett test (Ho: All correlation coefficients are not quite far from zero) is rejected on a level of statistical significance p<0.0001 for Approx. Chi Square=1159,108. Consequently, the coefficients are not all zero, so that the second acceptance of factor analysis is satisfied. As a result, both acceptances for the conduct of factor analysis are satisfied and we can proceed to it.

The scree test produces the following graph, which proceeds to a graphic representation of eigenvalues and guides us to the determination of the number of the essential factorial axes [14]. It presents a distinguished break up to the eighteen factors, whereas after the seventeen factors an almost linear part of the eigenvalue curve follows. Thus, we can take under consideration the eigenvalues.

Factor I

The 1st factor has an eigenvalue = 4,638. Since this is greater than 1.0, it explains more variance than a single variable, in fact 5,888times as much.

The percent a variance explained :

$$\left(4,638/17\text{ units of variance}\right)*\left(100\right)=27,281\%$$

Factor II

The 2nd factor has an eigenvalue = 2,141. It is also greater than 1.0, and therefore explains more variance than a single variable.

The percent a variance explained:

$$\left(2,141/17\text{ units of variance}\right)\left(100\right)=12,596\%$$

Factor III

The 3rd factor has an eigenvalue = 1,738. Like Factors I & II it is greater than 1.0, and therefore explains more variance than a single variable.

The percent a variance explained

$$\left(1,738/17\text{ units of variance}\right)\left(100\right)=10,222\%$$

Factors 4 through 17 have eigenvalues less than 1, and therefore explain less variance that a single variable.

The sum of the eigenvalues associated with each factor (component) sums to 17.

$$4,638+2,141+1,738+1,170+1,071+\ldots+0,146+0,070=17$$

Nota Bene

The data was suitable for the proposed statistical procedure of principal components analysis. Nine factors of annealing importance were derived to represent the data and were retained for further analysis [15].

On our observations, the results obtained three eigenvalues which contributed 50,099% of total variance; however, to achieve a minimum of 80%, eight eigenvalues were considered which explained 81,374% of the total variation. The expression of certain variables may be under the control of more than one factor, leading to an interaction between experimental factors. Only PCR are well adapted to cope with possible interactions; these interactions are identified because more than one factor plays a major role in the definition of an axis.

We have focused on interactions between categorical and continuous variables. However, there can also be interactions between two continuous variables. Suppose further it is believed that the effect of intentions on behavior (i.e. the correspondence between what one wants to do and what one actually does) is greater at higher levels of income. A positive value for the effect of the interaction term would

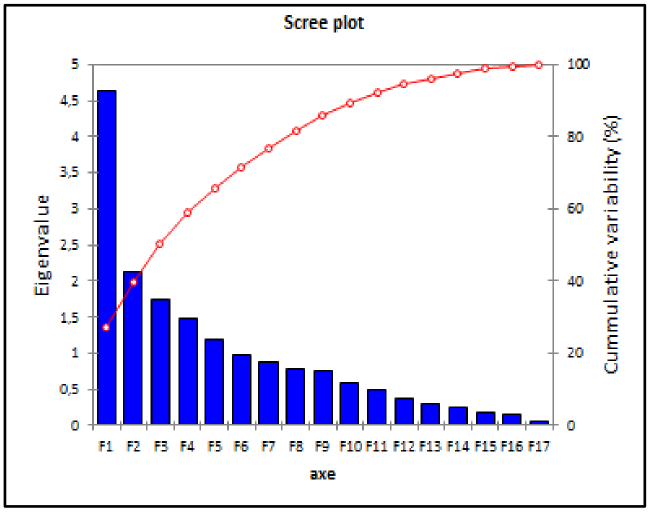

Fig 3: Scree plot

This method principle component in the regression lets we choose the best model from amongst all the models which can handle a number of variables varying from "Min variables" to "Max Variables". Furthermore, we can choose several "criteria" to determine the best model: Adjusted R², Mean Square of Errors (MSE), Mallows Cp, Akaike's AIC, Schwarz's SBC, Amemiya's PC.

It is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean of the squares of the errors (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

Computed against model Y=Mean(Y)

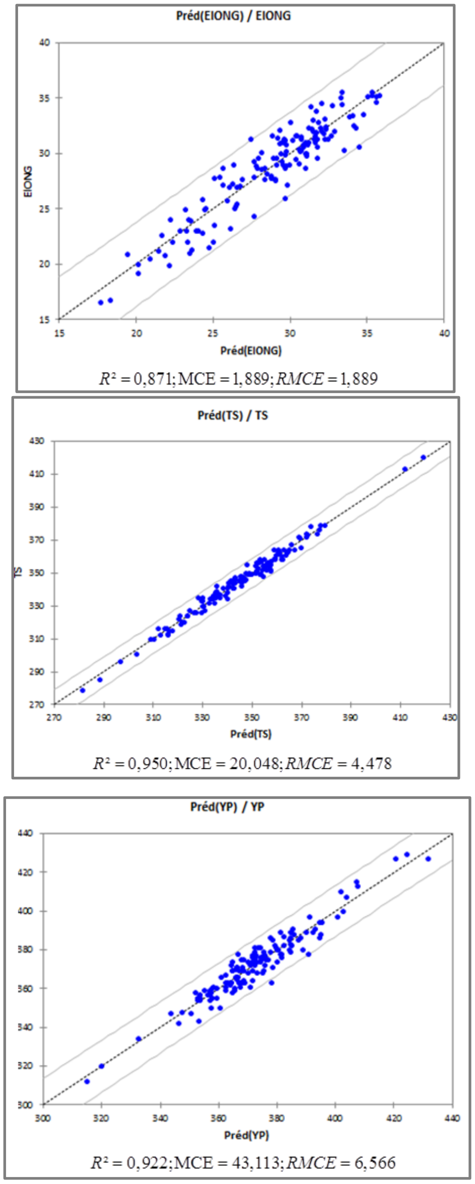

Fig 4: Results of principal component regression of the steel strip in an annealing heating furnace on hot-dip galvanazing line

When observing these results, performed with simple validation, a researcher might be tempted to use the models based on PCR

The fit of the model of the data can be evaluated statistically applying Analysis of variance (ANOVA), a residual analysis, or an external validation using a test set. One also can determine the significance of the b coefficients in the above model and then eliminate the non-significant ones, in practice, the optimum is not one point but a region with acceptable performance, the quadratic model without statistical performs acceptably well. The model is used to find the proper conditions and not for predictive purposes as multivariate calibration models are. Therefore less effort can be spent in finding best model and the quadratic one usually fits the data acceptably good [17].

The regression coefficient values of the shape conformity model. The p-values of each coefficient were used to examine the significance level, which also indicate the interaction effects between each independent variable. Whether a quadratic model is significant or not could be determined through ANOVA. As seen in table, ANOVA shows that this quadratic regression model is highly significant [18]

The predicted (modeled) values of primary productivity obtained using the regression model clearly coincided with the observed values (Fig. 8); this result proves the applicability of regression models in such studies. A univariate simple regression model has the advantage of yielding a high R² value. The present study confirmed that principal component regression analysis is useful for predicting complex processes (such as furnace on hot-dip galvanizing line) using environmental monitoring variables and understanding the relationships between parameters at this level with both, mechanical properties from one side and operational conditions from the other side.

5. Conclusion

This paper shows that the use of classic techniques of simple or cross validation for determining the best model based on historical data on the annealing process can lead us to choose models that closely fit products that have already been processed but which are less efficient when used for predicting new ones. In order to obtain over all prediction models that are capable of predicting the strip's dynamic performance in the event of temperature and speed fluctuations and which take in to account the size and type of steel on the coil being processed, it has been shown that PCR continue to be some of the more promising techniques for the design of overall prediction models and outperform other Data Mining techniques currently being used. The final model has proven to be efficient at dealing with new types of coils and process conditions. Its use can help to improve control systems and conveniently designed the parameters in transition zones between coils in order to achieve a more uniform treatment in this area.

In this work, only the variances of observed values were considered. Therefore, the variances of predicted responses can be another future research on this subject.

REFERENCES

[1] A. Sanz-García, J. Fernández-Ceniceros and F. Martínez-de Pisón, "A GA-SVR approach with feature selection and parameter optimization to obtain parsimonious solutions for predicting temperature settings in a continuous annealing furnace," *Applied Soft Computing*, vol. 35, pp. 13-28, 2015.

[2] F. Martínez-de Pisón, A. Pernía, E. Jiménez-Macías and R. Fernández, "Overall model of the dynamic behaviour of the Steel strip in an annealing heating furnace on a hot-dip galvanizing line," *revista de*

metalurgia, vol. 46, pp. 405-420, 2010.

[3] A. Sanz-García, F. Antojanzas-Torres, J. Fernández-Ceniceros and F. Martínez-de Pisón, "Over all models based on ensemble methods for predicting continuous annealing furnace temperature settings," *Ironmaking & Steelmaking,* vol. 41, no. 2, pp. 87-98, 2014.

[4] Y. Kim, K. Moon, B. Kang, C. Han and K. Chang, "Application of neural network to the supervisory control of a reheating furnace in the steel industry," *Control Engineering Practice,* vol. 6, no. 8, p. 1009–1014, 1998.

[5] Taha Hossein Hejazi, Mirmehdi Seyyed-Esfahani and Majid Ramezani, "New hybrid multivariate analysis approach to optimize multiple response surfaces considering correlations in both inputs and outputs," *Acta Scientiarum Technology*, vol. 36, no. 3, pp. 469-477 , 2014.

[6] B. Joaquín, J. Ordieres and M. Ordieres, "Data Miningin industrial processes," *Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005,* pp. 57-66 ISBN: 84-9732-449-8 © 2005 Los autores, Thomson, 2005.

[7] C. Théry, C. Biernacki et G. Loridant, «Model-Based Variable Decorrelation in Linear Regressio,» *HAL Id: hal-01099133* , 2014.

[8] T. Lundstedt , E. Seifert , L. Abramo , B. Thelin , A. Nystrom and J. Pettersen , "Experimental design and optimization," *Chemometrics and Intelligent Laboratory Systems,* vol. 42, pp. 3-40, 1998.

[9] D. MARLOW , "Mathematical modelling of the heat tratment in the continous processing of steel strip," *thesis from University of Wollongong,* 1995.

[10] M. Langer, N. Link , V. Torre , J. Ordieres Meré , E. Lindh-Ulmgren and M. Stolzenberg , "Investigation, modelling and control of the influence of the process route on steel strip technological parameters and coating appearance after hot dip galvanising," *Luxembourg: Publications Office of the European Union,* pp. 110 ISBN 978-92-79-12481-5, 2009.

[11] W. Glen, W. Dunn and D. Scott , "Principal Components Analysis and Partial Least Squares Regression," *Tetrahedron Computer Methodology,* vol. 2, no. 6, p. 349 to 376, 1989.

[12] P. Mujumdar, «Stochastic Hydrology,» Department of Civil Engineering Indian Institute of Science, Bangalore, 2011.

[13] S. Anastasiadou , «Reliability and validity testing of a new scale for mesuring attitudes toward learning statistics with techology,» *Acta Didactica Napocensia,* vol. 4, n° %11, 2011.

[14] E. BAIR, T. HASTIE, D. PAUL and R. TIBSHIRANI, "Prediction by Supervised Principal Components," *Journal of the American Statistical Association,* vol. 101, no. 473, 2006.

[15] T. Hejazi, M. Ramezani, M. Seyyed-Esfahani and A. Kimiagari, "Multiple Response Optimization with Probabilistic Covariates Using Simultaneous Equation Systems," *International Journal of Industrial Engineering & Production Research,* vol. 24, no. 2, pp. 113-121, 2013.

[16] R. Reris et P. Brooks , «Principal Component Analysis and Optimization,» *14th INFORMS Computing Society Conference Richmond, Virginia,* p. 212–225, 2015.

[17] A. Sandulyaka, A. Sandulyaka, F. B. Belgacemb and D. Kiseleva , "Special solutions for magnetic separation problems using force and energy conditions for ferro-particles capture," *Journal of Magnetism and Magnetic Materials,* vol. 10, no. 1016, pp. 902-906, 2016.

[18] A. Messaâdi, N. Dhouibi and F. B. Belgacem, "A New Equation Relating the Viscosity Arrhenius Temperature and the Activation Energy for Some Newtonian Classical Solvents," *Journal of Chemistry,* vol. 163262, p. 12, 2015.