Modeling and parameter optimization of the papermaking processes by using regression tree model and full factorial design

JOSÉ L. RODRIGUEZ-ALVAREZ, ROGELIO LOPEZ-HERRERA, IVÁN E. VILLALON-TURRUBIATES, GERARDO GRIJALVA-AVILA, AND JORGE L. GARCÍA ALCARAZ

ABSTRACT: One of the major challenges in the pulp and paper industry is taking advantage of the large amount of data generated through its processes in order to develop models for optimization purposes, mainly in the papermaking, where the current practice for solving optimization problems is the error-proofing method. First, the multiple linear regression technique is applied to find the variables that affect the output pressure controlling the gap of the paper sheet between the rod sizer and spooner sections, which is the main cause of paper breaks.

As a measure to determine the predictive capacity of the adjusted model, the coefficient of determination (R^2) and *s* values for the output pressure were considered, while the variance inflation factor was used to identify and eliminate the collinearity problem. Considering the same amount of data available by using machine learning, the regression tree was the best model based on the root mean square error (*RSME*) and R^2. To find the optimal operating conditions using the regression tree model as source of output pressure measurement, a full factorial design was developed. Using an alpha level of 5%, findings show that linear regression and the regression tree model found only four independent variables as significant; thus, the regression tree model demonstrated a clear advantage over the linear regression model alone by improving operating conditions and demonstrating less variability in output pressure. Furthermore, in the present work, it was demonstrated that the adjusted models with good predictive capacity can be used to design noninvasive experiments and obtain

Application: Through the strategy followed in this paper, the output pressure in the rod sizer could be controlled, as shown by the results. The approach used has not been exploited in papermaking processes.

Throughout history and until the end of the 20th century, it can be observed that the paper industry did not have the same growth as other industries, such as the automotive industry. However, global paper and paperboard production increased by 50% between 1990 and 2006 [1], thus giving the first signs of resurgence for this industry in some countries. The greatest increase in paper and paperboard production is in China, which ranges from 7% to 16% of global production, whereas in other countries production has fallen, including the United States, Japan, and Canada. However, since 2006, this industry has not experienced sustained growth due to less human consumption brought on by technological advances that have hindered this important industry, which has been a national emblem for many countries, such as Finland [2].

In 2018, the largest global producers of pulp were North America, Europe, Asia, and Latin America, while the most important paper producers were Europe, Asia, North America, and Latin America [3]. Thus, the economic situation for the pulp and paper industry located in traditional manufacturing countries has been uncertain for some time because of competition with countries having plentiful resources, low manufacturing costs, and large, newly established production facilities [4]. Currently, two new elements are revitalizing the industry in an unexpected way: 1) a boom related to e-commerce, which has notably increased the amount of paper and cardboard packaging throughout the world; and 2) environmental pressure to replace plastic packaging with more environmentally-friendly materials.

In Mexico, the Chamber of Paper highlights the importance of this sector, as it emphasizes that global production of the pulp and paper industry amounted to 400 million tons in 1999, against all the predictions for a decrease in paper consumption due to the boom of the electronic age [5]. In addition, technical advances and important investments in the last eight years have changed the production process, which now embraces methods that do not pollute, while also recycling paper that was normally discarded. Thus, the industry has contributed to developing and structuring the collection, production, and marketing activities for waste that is now considered a premium material in the papermaking process.

Subsequently, the pulp and paper industry in tradition-

al manufacturing countries has been forced to seek alternatives to be sustainable and profitable [1]. The linear regression technique has been widely exploited for this purpose in this type of manufacturing industry.

There are several documented applications of linear regression in different industrial sectors, fields of science, and specific areas. For instance, Belusso et al. [6] report use of a multiple linear regression model for pricing of Infrastructure-as-a Service, including the geographical location of the data center as one variable. Yazir and Sahin [7] propose a linear regression approach for a Black & Scholes model that is used for a call option and put option to derive pricing and risk management in the shipping industry. Meanwhile, for a global logistics system, Herrera and Park [8] proposed an early warning model based on principal component regression to predict a country's global logistics system risk, identify risk sources with probabilities, and suggest ways of risk mitigation. Muriyatmoko and Putra [9] focused their work in science and technology on how strong correlation impacts h-index towards citations using linear regression. These are just a few examples; according to Kumari and Yadav [10], the linear regression technique is widely used.

In many other applications, the regression models have been compared against other techniques. For instance, Abbas et al. [11] developed a mathematical model for predicting surface roughness from experimental data by using the regression analysis method, and the experimental data is then compared against the regression analysis and ANFIS (Adaptative Network-based Fuzzy Inference System) estimation. Gul and Guneri [12] developed two models (regression and neural network) for forecasting patient arrivals at emergency departments with short- and long-term plans, as well as integration of physical capacity requirements, staffing, budgeting, and arranging staff schedules. While Tsai and Tsai [13] presented an industrial application of artificial neural network (ANN) and support vector regression (SVR) to diagnose the control reflow soldering process in a closed-loop framework.

In applications related to the paper industry, Adamopoulos et al. [14] presented a predictive model for the mechanical properties of corrugated base papers (liner and fluting medium) from fiber and physical property data using multiple linear regression and artificial neural networks. Kilulya et al. [15] used a partial least squares (PLS) regression model to evaluate the effects and influence of the lipophilic extractive residues on quality parameters of dissolving pulp, and their findings indicate that sterols, fatty alcohol, saturated and unsaturated fatty acids significantly influenced/affected viscosity, kappa number, and carbohydrates in the pulp. Meanwhile, Marklund et al. [16] modeled the influence of fiber properties on strength parameters for softwood kraft pulps made from 20 different types of wood samples by using multivariate data analysis and partial least squares method.

Recently, the use of regression trees has been applied in many fields of science, such as smart energy efficient buildings systems [17], construction injury prediction [18], wastewater treatment plant performance modeling [19], landslide susceptibility studies [20], and data mining [21], to mention a few. Similarly, in other specific applications, scientists have used regression trees to predict I-V characteristic curves; create a flood susceptibility map; assess the importance of plant, soil, and management factors affecting potential milk production on organic pastures; assess a wastewater treatment facility's compliance with decreasing ammonia discharge limits; estimate mass first flush ratio in urban catchments; analyze forest stand damage susceptibility during harvesting; determine the relationship between meteorological factors and mumps; predict overweight trucks from historical weigh-in motion data; and forecast upland rice yield under climate change [22-30].

The theoretical foundations and practical applications of classification and regression trees were first presented by Breiman et al. [31]. Classification and regression trees are local regression models that work in a recursive manner to partition the predictor space by delineating regions where the predictions of a dependent or response variable to a set of predictors is homogeneous. These trees first group the predictors, and then a particular model is assigned to each of these groupings, with the simple average most commonly used [32]. Local models have the advantage of representing the true relationship between the covariates and the response variable in each of the groups in the tree, as compared to global models where a single equation is used for the entire data set. If the response variable is continuous, then it is called a regression tree; if categorical, then it is called a classification tree [27].

The interpretability of the tree structure is a strong reason for the regression tree's popularity among the practitioners, along with good prediction accuracy, fast computation speed, and wide abilities. Moreover, a clear advantage of regression trees is that they make no probability distribution assumption; however, they do identify relevant explanatory variables (or features) and detect interactions among them [23].

For years, theory and designed experiments applications were consolidated. In several industries, the contributions of Douglas C. Montgomery and Genichi Taguchi paved the way for routine applications, with this approach being one of the strategies that companies have used in their improvement projects. To comply with a quality characteristic when changes are made within processes, trial and error are usually an alternative. On the other hand, the design of experiments (DOE) approach uses all possible combinations of variables, which are analyzed by using factorial design [33]. Finally, the regression tree model is used to describe the response variable and conduct a full factorial DOE in order to find the optimal operating parameters.

Although use of linear regression and regression trees

1. Rod sizer-spooner section of a papermaking machine.

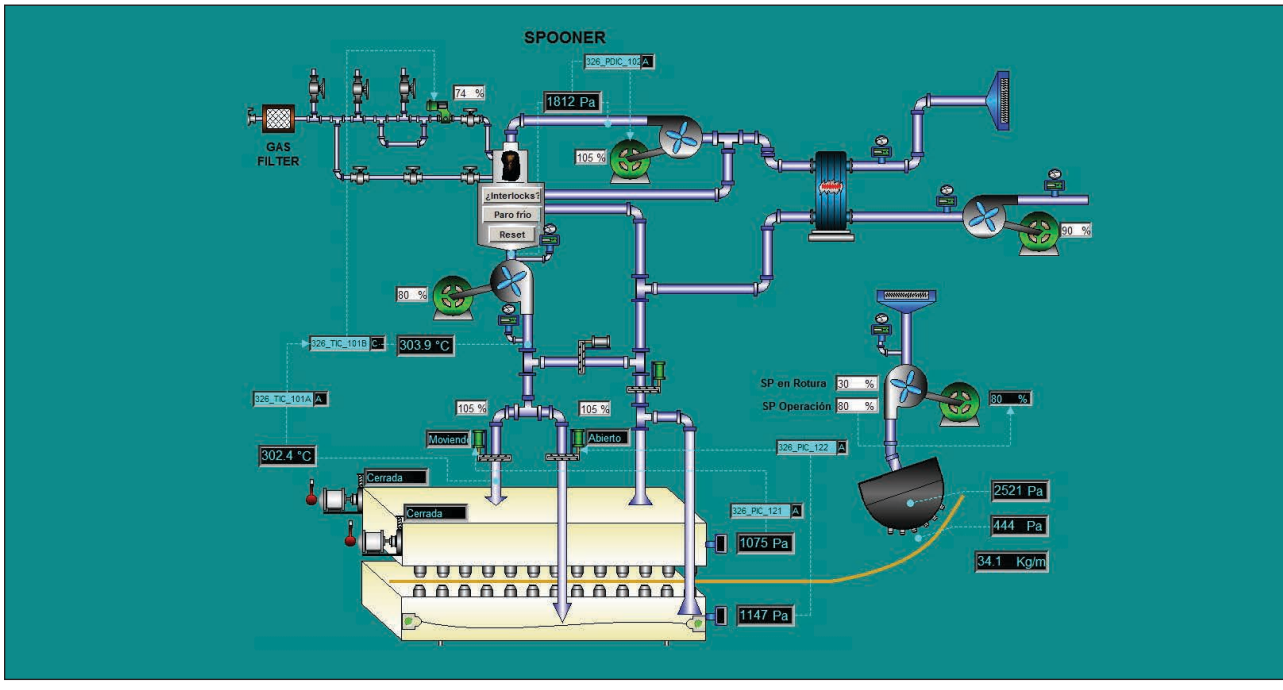
through application of machine learning techniques has been widely adopted for various problems in the pulp and paper industry, machine learning has not been exploited in the rod sizer and spooner section of the papermaking process.

PROBLEM AND RESEARCH OBJECTIVE

Although linear regression techniques have been widely used in the paper industry, they have not been applied in the rod sizer and spooner section of a paper machine that uses only recycled raw material until now.

One of the common problems in paper manufacturing is paper breaks that occur in different machine sections and mainly between the rod sizer and spooner section, as shown in **Fig. 1**. The path of the paper in this section is controlled through an air box, as shown in **Fig. 2**. The air output pressure must be stable to avoid the paper experiencing a sudden change in tension that causes the paper to break. Currently, paper breaks in the rod sizer and spooner section represent an average of 349 min/month of machine downtime. Most breaks occur when the paper undergoes aggressive changes in tension due to the variation in air output pressure (pressure greater than 100 Pa). Thus, the aim of this study is to reduce the process variation below 60 Pa, using this value as a safety factor. In order to find the correct values to reach an air output pressure of 60 Pa, the desirability index will be used to visualize the variability reached after setting the new values in the process.

In summary, the present research approach considers use of the machine learning technique to develop predictive models as a first step and, ultimately, the use of designed experiments to obtain optimal operating parameters. By using this approach, it is possible to reduce variability in the air output pressure, in addition to reducing economic losses from machine downtime related to paper breaks between the rod sizer and spooner section. Furthermore, if a predictive model can be developed, then a DOE can be carried out without an invasive process, which means that the process does not need to stop due to responses in the experiment that are calculated by using the predictive model previously developed. This research approach is useful in papermaking processes if the three following conditions are fulfilled: a large amount of data is available; input and output variables are fully identified: and the process staff have sufficient technical skill level.



2. Air box (air tum).

MATERIALS AND METHODS

The present paper attempts to contrast the use of the regression tree as an alternative to the linear regression technique, with the goal of evaluating the results and finding the best solution (optimal parameters) for eliminating the problem of paper breaks. This section details the process that was followed, from data collection, models development, and DOE to finding the optimal parameters.

Data collection

The first task was to identify the main variables (input and output) that affect paper breaks. **Table I** illustrates variable names (like those shown on the control screen), units, and codes used for modeling. As a second step, data collection was structured and debugged. The data used correspond to a one-month period and cover all possible operating conditions. The data were extracted from the server. To develop a model sensitive to minor changes, data were extracted at 1-min intervals, and a total of 40,333 readings were initially obtained. Subsequently, after data debugging, some data that were atypical from input and output values were discarded, leaving a total of 35,971 readings [34]. This amount of collected data was used to generate the linear regression model.

Regression model

A linear regression model (LRM) attempts to model the relationship between two or more explanatory variables, as well as a response variable, by fitting a linear equation to observe data [11]. Considering $x_1, x_2,\ldots,x_n$ as a set of n independent variables (estimators) associated with a value of the dependent variable y, the linear regression model for the j^{th} sample is obtained according to Eq. 1:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_n x_{jn} + \varepsilon \quad (1)$$

where ε is a random error and β_i (i=0,1,2,…,n) are the unknown regression parameters. A multiple LRM was fitted by using Minitab software (Minitab; State College, PA, USA).

The least squares technique was the method used for estimating the parameters. Least squares is a method that adds the squares of the residuals or errors around the regression line to minimize them [35]. As suggested by Pulido et al. [36], the models are considered highly reliable, as during the sampling process, data do not violate the assumptions of normality, independence, and constant variance. The determination coefficient (R^2) is the considered index to select the best model. This index is interpreted as the proportion of total variation in the y that is explained by the x_i variables [37]. Since, this research reports a multiple linear regression model, the adjusted determination coefficient R^2_{aj} is therefore used. This coefficient explains the variability of the model [38], and it is calculated by using Eq. 2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i')^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (2)$$

where Y_i is a measured value, Y_i' is a predictive value, $\overline{Y}$ is the mean of the measured values, and the number of instances in a test set is given by n. The resulting variance analysis considers a significance value of 5%.

In addition, it was necessary to test collinearity among independent variables. The collinearity problem deals with linear relationships between two or more independent variables. In a linear model, as shown in Eq. 3 with n observations and p independent variables, it is assumed that it is correctly specified, that the model has mean zero (there is no y-intercept parameter β_0), and that there is homoscedasticity and uncorrelated error:

$$Y = X\beta + \mu \quad (3)$$

Variable Name	Variable Name on Control Screen	Coding
Air output pressure	Output Press	Y
Differential pressure in dryer 50B	Reg Pres Sec 50B	x_1
Differential pressure in dryer 50	Reg Pres Sec 50	x_2
Differential pressure in dryer 52 54 56	Reg Pres Sec 52 54 56	x_3
Pressure in dryer 50 51	Pres Sec 50 51	x_4
Pressure in dryer 52 54 56	Pres Sec 52 54 56	x_5
Pressure in dryer 53 55 57	Pres Sec 53 55 57	x_6
Steam pressure Line 4	Pres Steam 4 bar	x_7
Steam pressure Line 10	Pres Steam 10 bar	x_8
Internal pressure of air box	Int Pres	x_9

I. Variables in the model.

If the collinearity is perfect, the correlation coefficient between two independent variables is 1, and thus, there is no unique solution or the estimations will be unstable [39]. In order to solve the collinearity problem, some authors have suggested the Ridge regression technique [40], the surrogate Ridge model [41], and nested regression [42], among others. A recent approach to deal with this problem was developed by Piña [43].

On the other hand, mechanically removing variables can be used to deal with heteroscedasticity and autocorrelation. In the present work, this approach has been considered for treating the collinearity problem: in the meantime, the variance inflation factor (VIF) is used to measure it. The VIF index allows measurement of collinearity's impact for the variables x_i, $i=1,\ldots,p$ with the rest of the independent variables [44]. VIF is calculated as shown in Eq. 4:

$$VIF_{(i)} = \frac{1}{1 - R_i^2}, \quad i = 1, \ldots, p \tag{4}$$

where R_i^2 is the coefficient of determination of x_i on the rest of the independent variables. Although multicollinearity should not be a major concern when fitting models with large sets of data, those models should be used for predictive purposes. A VIF of 5 is acceptable [45]. In the present research only, variables that have a VIF under 2 were considered [44] due to the precision that is required in the paper production process.

When a predictive model is defined, then it is possible to find the operating parameters that allow reducing the variability in air output pressure. This is the main objective and a common interest in all improvements that involve the determination of the optimal setting conditions for independent input variables and to attain the best solution for dependent responses [46]. A frequent practice is to find the optimal setting condition by transforming the response to a desirability value. Harrington proposed the first original transformation of a response [47]. The desirability index function converts all the desirability values, d_j's, into a single composite desirability.

The composite desirability is maximized by conventional or unconventional searching techniques. The composite desirability value (D) always lies between 0 and 1. The equation developed by Harrington in 1965 combines the individual desirability with a composite desirability (D), as shown in Eq. 5:

$$D = maximize\left(\prod_{j=1}^{r} d_j\right)^{\frac{1}{r}} \tag{5}$$

where d_j is the individual desirability value.

Regression trees

As an alternative to the traditional linear regression approach, the supervised machine learning technique is used in the present work for the purpose of training the collected data and trying to find a better model than linear regression. Due to the fact that the rod sizer section processes all grades of paper, an alternative model can be an excellent option, in addition to the advantages of not assuming probability distributions and maintaining the ability to identify characteristics and interactions between them.

The classification regression trees approach is a heuristic trees method that unpacks the relationships between an outcome measure and a group of predictors. In the terminology of the classification regression tree, each box represents a group or sub-group called a node, and the node on the top of a tree is called the root node because the analysis descends from it. The classification regression tree analysis is full of partitions at different branches or at different levels. A partition refers to the splitting of cases in a node into groups. When a partition is made in a classification regression tree, one node produces two resulting nodes. The nodes produced are called child nodes, whereas the producing node is called the parent node. The node that cannot be further partitioned into child nodes marks the end of growth in that part of the tree, and it is called a terminal node [48].

For classification regression trees, the partitioning of cases into groups at each level is guided not by any statistical test but by a statistical criterion referred to as impurity [31]. Impurity measures the degree to which cases in a group belong to different categories (values) of the dependent variable. Although impurity can be conceptually defined in different ways, all measures of impurity share the same behavior. The impurity of a node is zero if all cases in that node belong to a single category of the dependent variable, and impurity becomes large if an equal number of cases belong to different categories of the dependent variable. One popular impurity measure is the entropy impurity given by Eq. 6:

$$i(\tau) = -\sum P(c_j) \log P(c_j) \tag{6}$$

where $P(c_j)$ represents the probability that a case falls into the category c_j or the proportion of the cases that go into that category in node τ. Logarithm is base 2.

A closely related issue to this discussion is when one should stop partitioning [48]. A stopping rule is used to stop the partitioning process. Caution is needed when setting the stopping rule. If the rule stops the partitioning too soon, the resulting tree is likely to be too small to reflect the structure of the data. In other words, the error in the structure of the tree tends to be large, which compromises the function of the tree. If the rule stops the partitioning too late, the resulting tree is likely to be too large to be either stable or meaningful. In other words, the tree becomes practically useless even though the error in the structure of the tree tends to be small.

Term	Coef.	SE Coef.	T-Value	p-Value	VIF
Constant	-1505.6	43.0	-35.05	0.00	
Reg Pres Sec 50B	-373.3	14.9	-25.05	0.00	167.70
Reg Pres Sec 50	573.0	15.5	36.94	0.00	173.31
Reg Pres Sec 52 54 56	968.8	12.3	78.46	0.00	5.66
Pres Sec 52 54 56	-605.5	17.9	-33.83	0.00	12.23
Pres Sec 53 55 57	-61.4	13.1	-4.71	0.00	6.94
Pres Steam 4 bar	26.05	7.63	3.41	0.00	3.70
Pres Steam 10 bar	9.222	0.628	14.68	0.00	1.13
Int Pres (Pa)	1.80621	0.00494	365.68	0.00	1.63

SE Coef. = standard error coefficient; VIF = variance inflation factor.

II. Initial coefficients and p-values.

There are several different ways to set the stopping rule. Traditionally, one adopts the notion of hypothesis testing to decide when to stop the tree [49].

In a classification problem, we have a training sample of n observations on a class variable Y that takes values 1,2,…,k, and p predictor variables, X_1,… ,X_p. Our goal is to find a model for predicting the values of Y from new X values. In theory, the solution is simply a partition of the X space into k disjoint sets, A_1,A_2,… ,A_k, such that the predicted value of Y is j if X belongs to A_j, for j=1,2,… ,k [50]. The steps followed for classification trees construction are:

1. Start at the root node.
2. For each X, find the set S that minimizes the sum of the ode impurities in the two child nodes and choose the split $\{X^* \in S^*\}$ that gives the minimum overall X and S.
3. If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn.

Although regression trees can be generated by using many available software applications, the regression trees in the present work were obtained using the trademarked Matlab R2019 software (Mathworks; Natick, MA, USA) and using a holdout validation technique that is recommended for a large data set. For training, the model used 80% of collected data and 20% was held out to validate the model.

Full factorial design

It is widely accepted that the most commonly used experimental designs in manufacturing companies are full and fractional factorial designs at two and three levels. Factorial designs would enable an experimenter to study the joint effect of factors (or process/design parameters) on response [51]. A full factorial designed experiment consists of all possible combinations of levels for all factors. The total number of experiments for studying k factors at 2-levels is 2^k. In order to find the optimal operating conditions by using the trained regression tree model, a full factorial design 2^9 is developed.

The regression tree model that was previously trained was used to calculate the air output pressure (response variable) for each experiment. The results were to be analyzed considering a significance value of 5%. The stepwise method was used to remove the input variables with an alpha value of 0.15 when entering and removing variables.

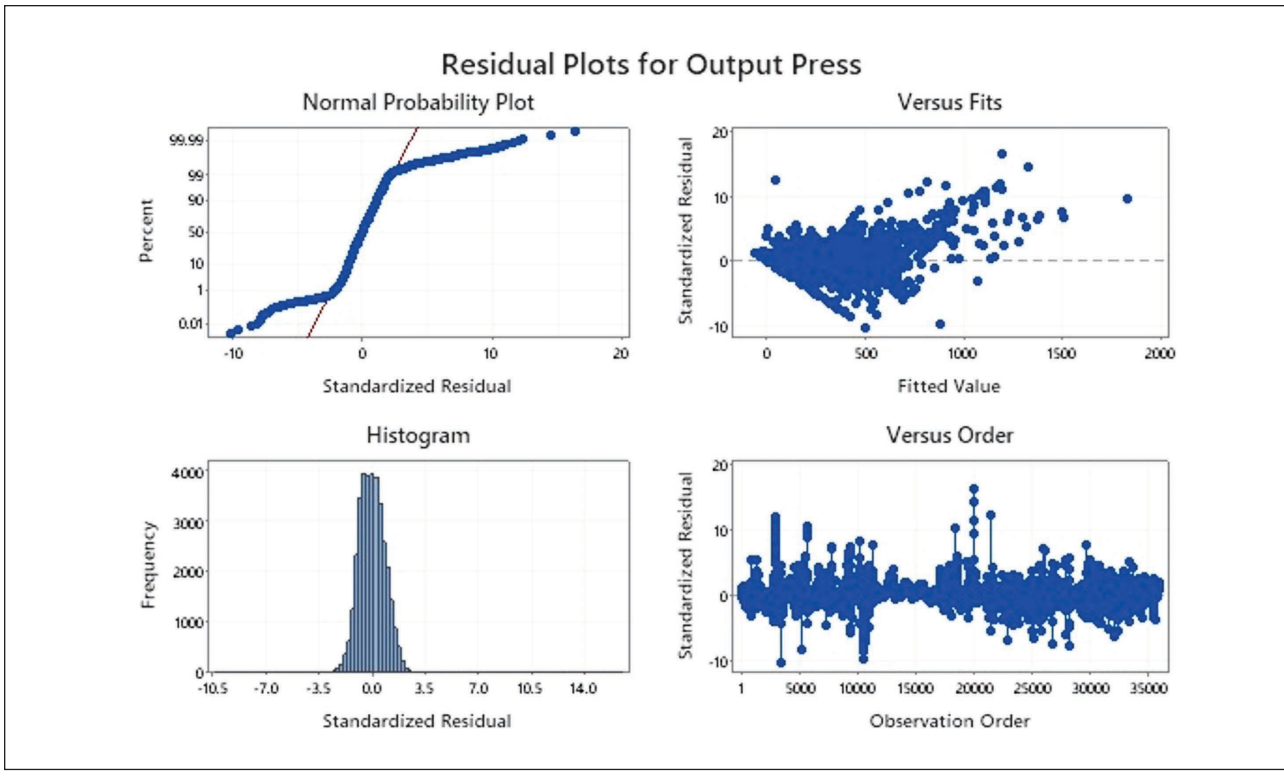
RESULTS AND DISCUSSION

After debugging the collected data, the aim of this paper was to fit a multiple linear regression model by considering all the variables shown in Table I. The data was processed by using Minitab software. The results in **Table II** show that the constant term and only eight variables were statistically significant. The pressure of the 50 51 (Press Sec 50 51) dryers were not significant. The resulting regression equation is presented in Eq. 7:

$$y = -150.60 - 373.30\,x_1 + 573.00\,x_2 + 968.80\,x_3 - 605.50\,x_5 - 61.40\,x_6 + 26.05\,x_7 + 9.22\,x_8 + 1.80\,x_9 \quad (7)$$

For the model shown in Eq. 7, the results shown a determination coefficient of 0.8529 (85.29%) and a standard deviation (s) of 44.6 Pa. **Figure 3** shows that normality, constant variance, and independence of the residuals are fulfilled.

The results in Table II shown two variables with the highest VIF values, indicating a serious collinearity problem. Thus, the Reg Pres 50B and Reg Pres 50 were removed from the model. In addition, because the Steam Press 4 bar does not directly impact the rod sizer and spooner section, this variable was also removed. After removing these independent variables, the Reg Pres Sec 52 54 56 and Pres Sec



3. Assumptions of the model.

52 54 56 variables showed the higher VIF in the following analysis; however, because Reg Pres Sec 52 54 56 is easier to control in process than the Pres Sec 52 54 56 variable, Pres Sec 52 54 56 was removed. The resulting independent variables are shown in **Table III**, and the better model is shown in Eq. 8:

$$y = -2456.40 + 253.03\,x3 - 180.64x_6 + 10.41\,x_8 + 1.65\,x_9 \quad (8)$$

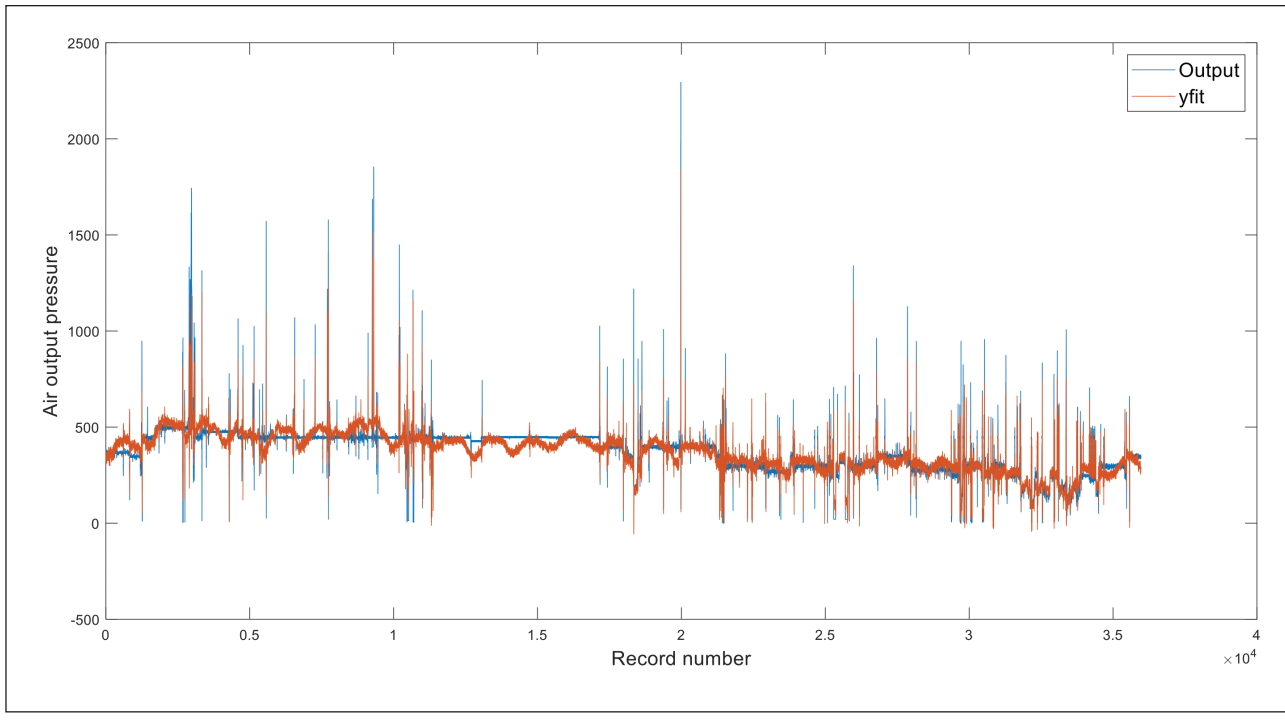
With this model, the results shown a determination coefficient of 0.8302, indicating that the model explains 83.02% of the total variability in air output pressure. In addition, the predicted standard deviation is 47.97 Pa. Although these results are a bit more adverse than the results shown by the model in Eq. 7, this model eliminates the collinearity problem.

By using the model shown in Eq. 8, a comparison between the observed and predicted values is presented in **Fig. 4**. This model can predict the air output pressure with good precision, and then the model can be used to find the optimal values to keep the process in control (rod sizer-spooner section).

To calculate the composite desirability index (D), the response optimizer tool of Minitab 19 was used. **Figure 5** shows a *D-index* value of 1.000, and the optimal conditions are as follows: differential pressure in dryer 52 54 56 (Reg Pres Sec 52 54 56) was 0.44 bar; pressure in dryer 53 55 57 (Pres Sec 53 55 57) was 4.23 bar; steam pressure in line 10 (Pres Steam 10 bar) was 9.71 bar; and an internal pressure in

Term	Coef.	SE Coef.	T-Value	p-Value	VIF
Constant	-2456.4	30.0	-81.98	0.00	
Reg Pres Sec 52 54 56	253.03	6.08	41.60	0.00	1.19
Pres Sec 53 55 57	-180.64	5.63	-32.09	0.00	1.12
Pres Steam 10 bar	10.411	0.667	15.62	0.00	1.10
Int Pres (Pa)	1.65272	0.00451	366.49	0.00	1.18
SE Coef. = standard error coefficient; VIF = variance inflation factor.					

III. Final coefficients and p-values.



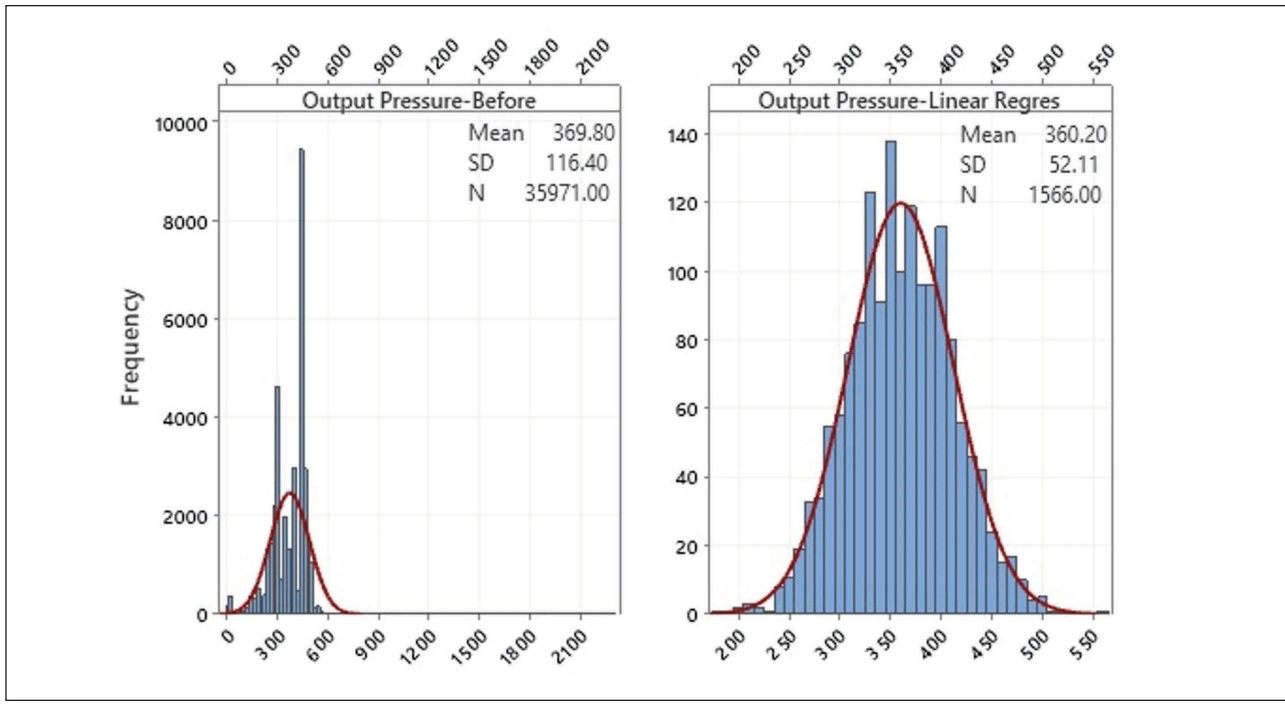
4. Observed output pressure vs. y-fit linear regression model.

5. Composite desirability and the optimal setting condition.

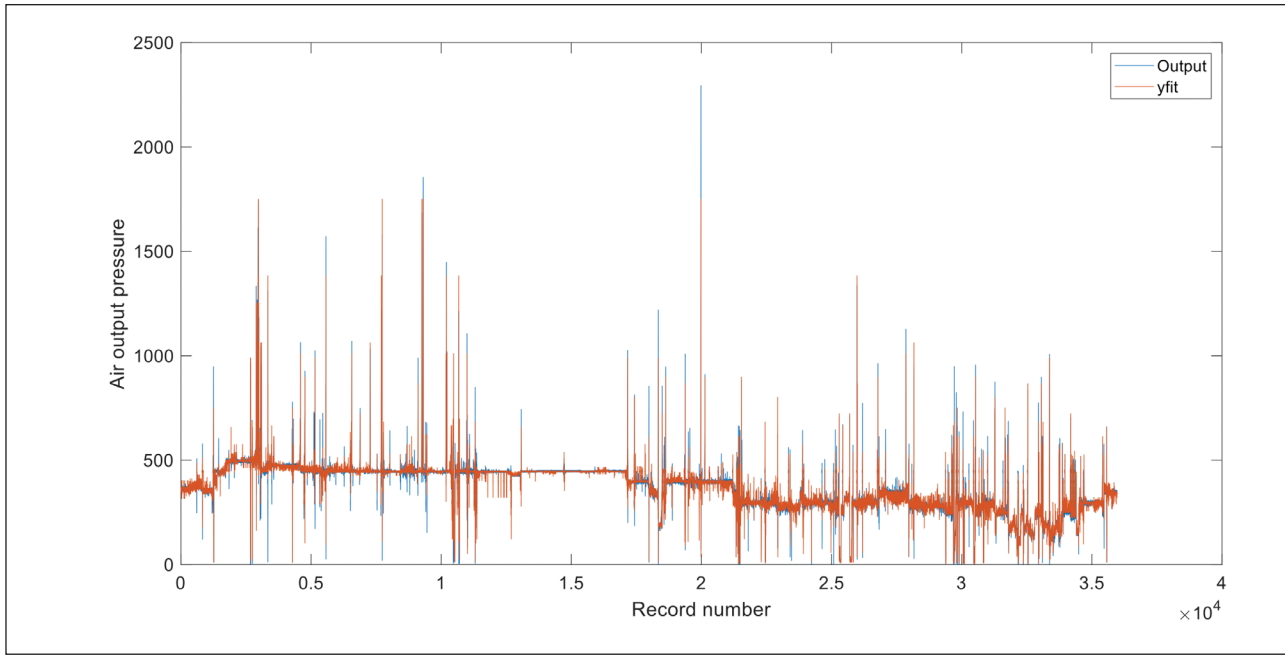
air box (Int pres) was 2038 Pa. With these operating parameters along with a 95% confidence level, the air output pressure was 359.99 Pa. The expected variation is from 350.80 to 369.18 Pa. After the optimal values were set in the process, the variability in the air output pressure markedly decreased. **Figure 6** shows a comparison between the process before and after setting the optimal values.

As an alternative, a regression tree model was obtained by using the machine-learning application in Matlab R2019 with the same collected data (35,971). The determination coefficient (R^2) and root square mean error (RMSE) were the criteria used for selecting the best model. The results shown an R^2 of 92% and RMSE of 32.50. The regression tree model has very good ability to predict the air output pressure, as shown in **Fig. 7**.

Because the previously-trained regression tree model has good ability to predict the air output pressure, then it was used to evaluate the response in the designed experiment. The approach used is a full factorial 2^9 design. All independent variables were coded, and their levels (low and high) were defined as shown in **Table V**. The levels of the variables were defined both based on the process engineers' experience and based on optimal parameters found by the previously-analyzed linear regression model.



6. Observed output pressure before and after setting the parameters found by linear regression (SD = standard deviation; N = sample size).



7. Observed output pressure vs. y-fit medium regression tree model.

In order to develop the designed experiment, a matrix of 512x9 size was first generated in Minitab. The response for each experiment was calculated by using the *yfit=c.predictFcn(X)* function in Matlab.

The results show an acceptable standard deviation (*s*) of 7.87, while the variability explained by the model is also acceptable, with an adjusted determination coefficient (R^2_{adj}) of 96.91%. The individual effects of differential pressure in dryer 52 54 56 (C), pressure in dryer 50 51 (D), steam pressure in line 10 (H), and internal pressure of the air box (J) were significant, as shown in **Table VI**. Meanwhile, the two-way interactions of differential pressure in dryer 52 54 56 (C)*pressure in dryer 50 51 (D); differential pressure in dryer 52 54 56 (C)*steam pressure line 10 (H); pressure in dryer 50 51 (D)*steam pressure line 10 (H); pressure in dryer 50 51 (D)*internal pressure of the air box (J); and

Variable Name	Code	Levels	
		Low (-)	High (+)
Differential pressure in dryer 50B	A	3.50	4.50
Differential pressure in dryer 50	B	3.50	4.50
Differential pressure in dryer 52 54 56	C	0.35	0.45
Pressure in dryer 50 51	D	3.50	4.50
Pressure in dryer 52 54 56	E	4.00	4.50
Pressure in dryer 53 55 57	F	4.00	4.50
Steam pressure Line 4	G	4.00	5.00
Steam pressure Line 10	H	9.00	9.50
Internal pressure of air box	J	2000.00	2050.00

V. Code and levels for each input variable.

Source	DF	Adj SS	Adj MS	F-Value	p-Value
Model	9	994893	110544	1781.61	0.000
Linear	4	935765	233941	3770.39	0.000
C	1	206972	206972	3335.74	0.000
D	1	28620	28620	461.27	0.000
H	1	629	629	10.14	0.002
J	1	699543	699543	11274.41	0.000
2-Way Interactions	5	59128	11826	190.59	0.000
C*D	1	28620	28620	461.27	0.000
C*H	1	629	629	10.14	0.002
D*H	1	629	629	10.14	0.002
D*J	1	28620	28620	461.27	0.000
H*J	1	629	629	10.14	0.002
Error	502	31148	62		
Total	511	1026041			

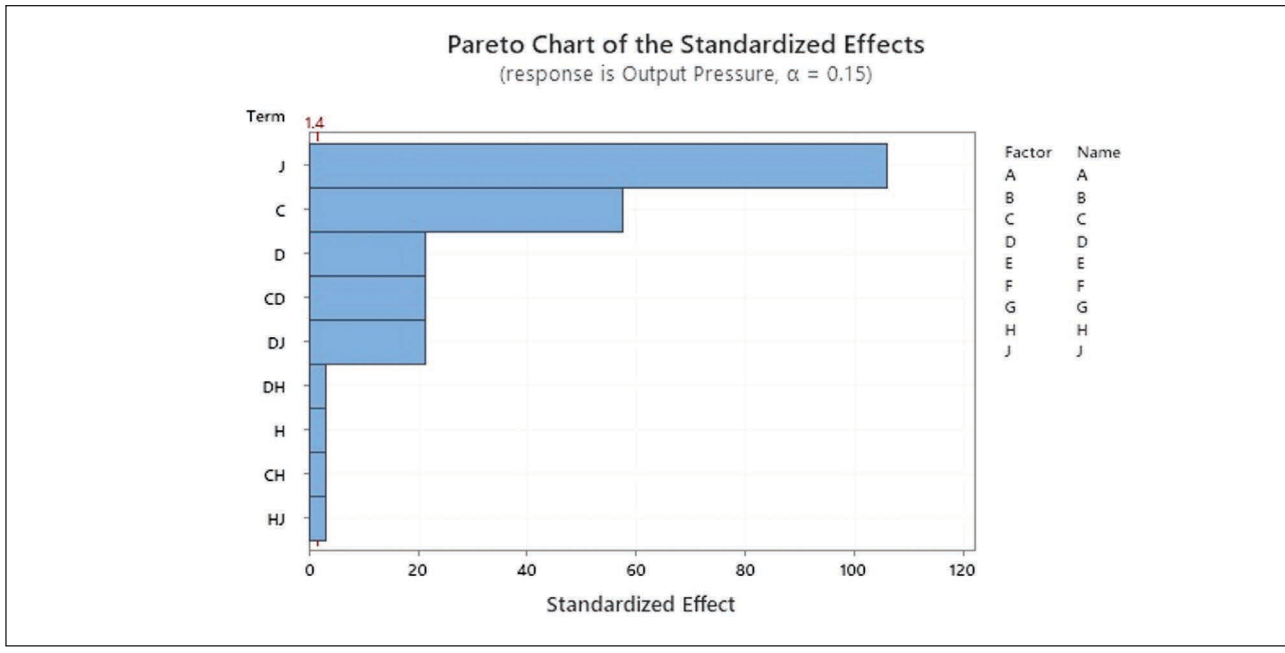
DF = degree of freedom; Adj SS = adjusted sum of squares; Adj MS = adjusted mean squares.

VI. Analysis of variance.

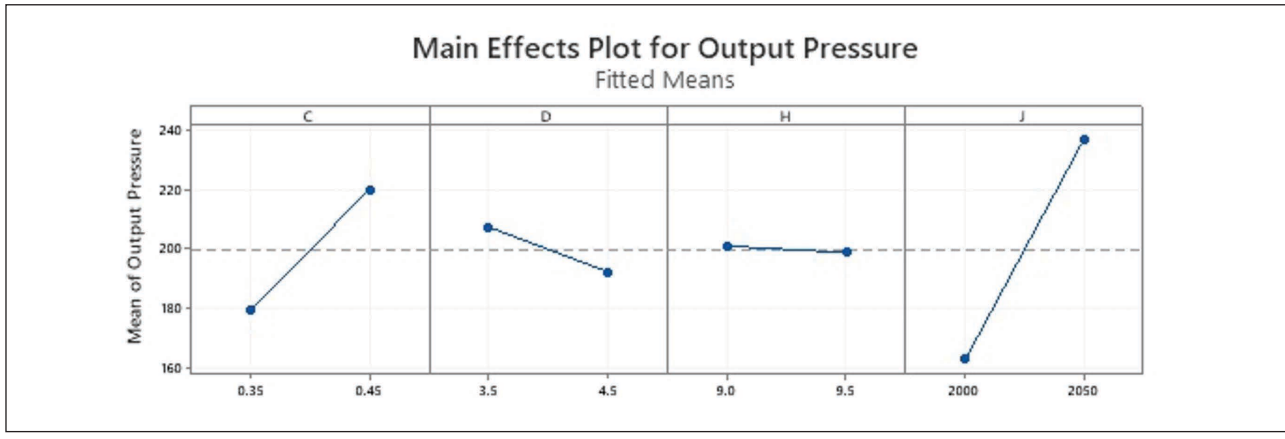
steam pressure line 10 (H)*internal pressure of the air box (J) were significant, as shown in Table VI. **Figure 8** shows the magnitude and importance of the individual effects and interactions.

The optimal operating parameters were obtained from the main effects and interactions plots, as shown in **Fig. 9** and **Fig. 10**, respectively. Since the interactions were significant, the optimal operating conditions are obtained from the interaction plot. Thus, from Fig. 10, the optimal operating conditions are 0.45, 3.50, 9.00 and 2050.00 for differential pressure in dryer 52 54 56 (C), pressure in dryer 50 51 (D), steam pressure line 10 (H), and the internal pressure of the air box (J), respectively. The rest of the input variables were removed because they had no effect on the air output pressure.

A lower air output pressure indicates that the paper web will be closer to the air box (air tum shown in Fig. 2). The optimal operating conditions found by using the regression tree model and designed experiment approach were tested over 10 h. In addition, **Fig. 11** shows a comparison between linear regression and the regression tree model developed. The results shows that the regression tree model has the lowest values in the mean and standard deviation, demonstrating a clear advantage over the linear regression



8. Pareto chart of the magnitude and importance of individual effects and interactions.



9. Main effects plot for response.

model. Finally, by using the records that were used in the process, a time-based bar plot for the paper breaks per month is presented in **Fig. 12**. The results shown that no paper breaks were observed over a three-month period, while the output pressure was below 100 Pa.

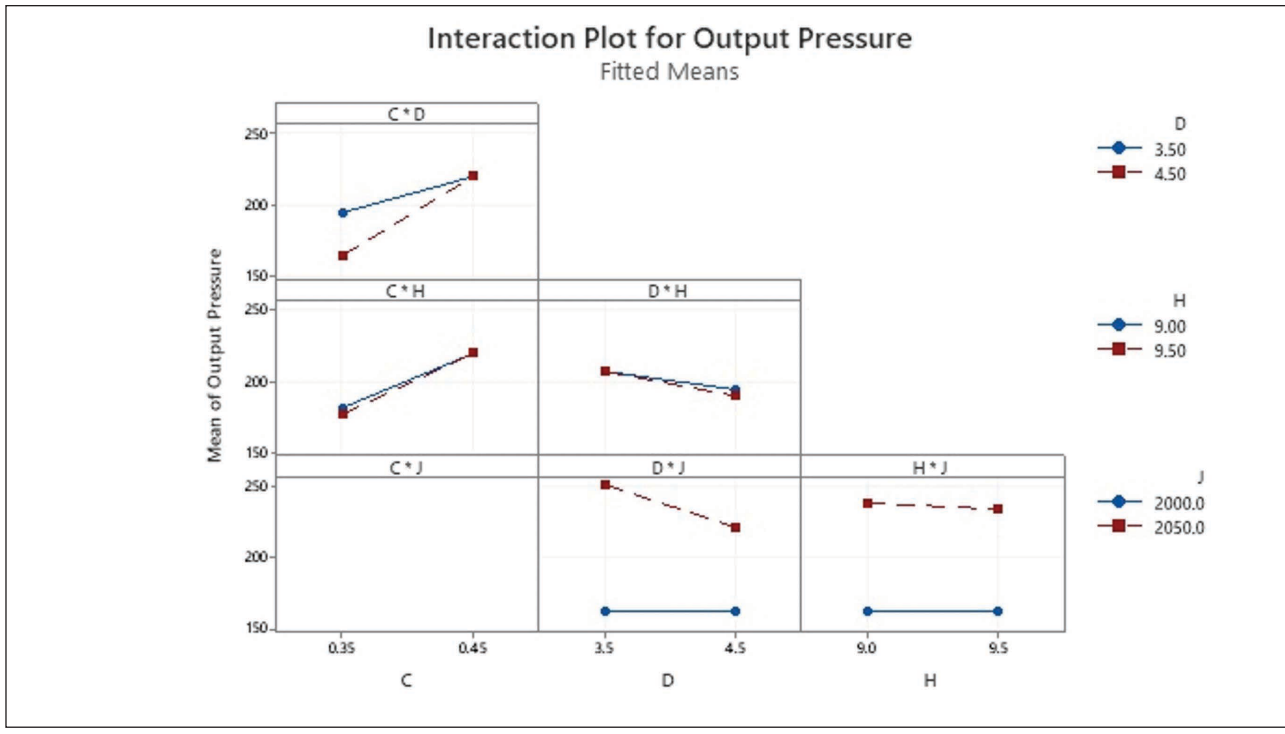
Although the optimal operating conditions found by the two methods are very similar, by using the regression tree model, the process was able to work with a lower pressure in line 10, generating a saving of 0.71 bar.

CONCLUSIONS

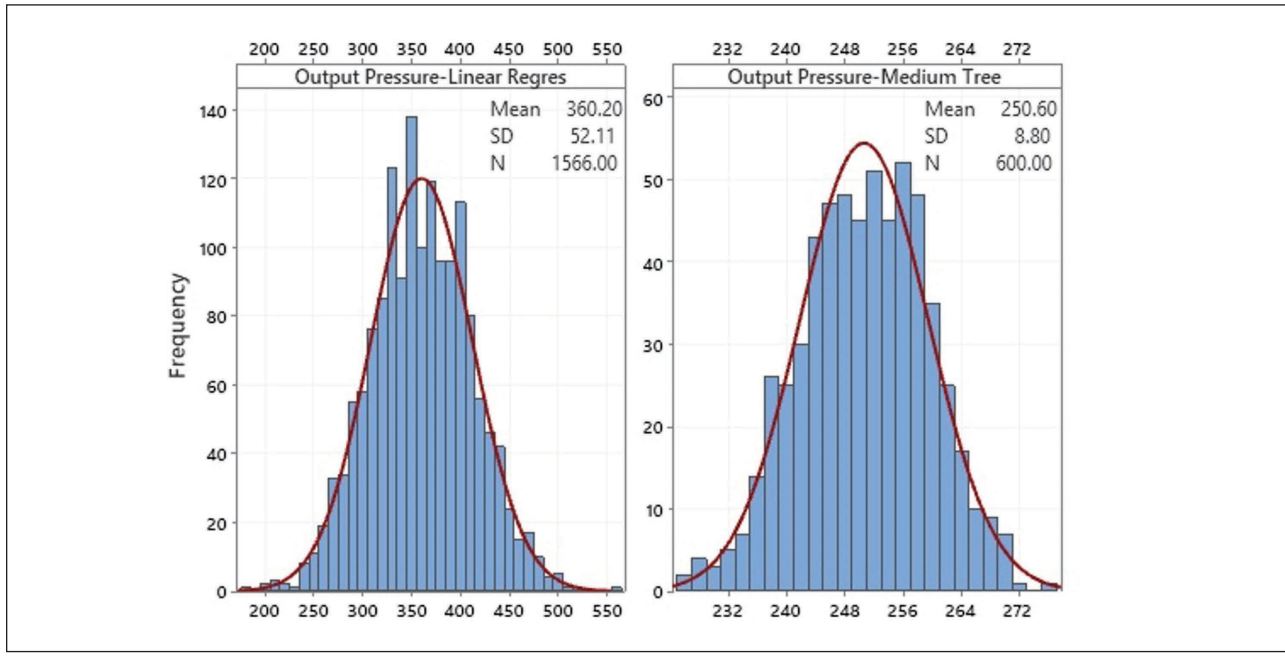
The instability in the air output pressure of the air box through which the paper passes in the rod sizer of a paper machine is the main cause of most paper breaks and resulting downtime in this section. This specific papermaking section has not been exploited by using regression techniques.

Taking advantage of available data as a first step, a linear regression model was trained and developed. This research demonstrated that fitted models have an acceptable predictive capacity, and the estimated operating parameters allowed a reduction in variability in the air output pressure, and therefore, minimization of paper breaks. Although the collinearity problem was not the purpose of this work, the mechanical solution for this problem is an excellent alternative when the process is widely known. After three months, the optimal operating conditions found by the regression model were acceptable. While the air output pressure was below 100 Pa, no stoppages due to variability problems were reported, as shown in Fig. 12.

As alternative using the same collected data, a regression model was trained by using the machine learning application in MatLab. Based on R^2 and *RMSE*, the medium regression tree was the best model. The results show that the



10. Interaction plot for response.

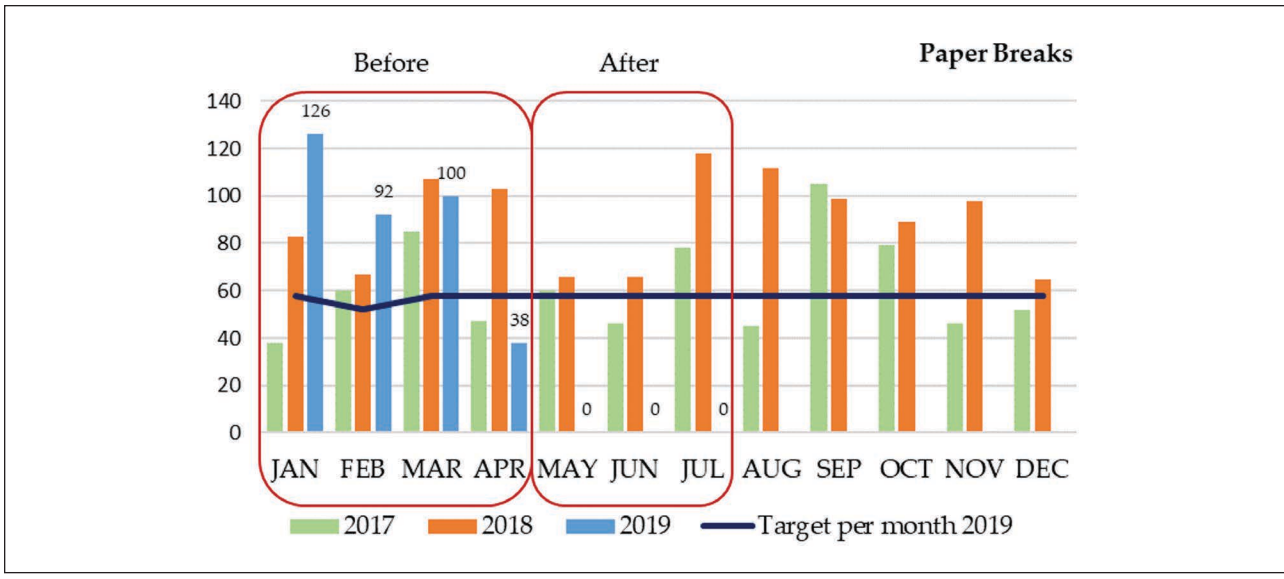


11. Observed output pressure before and after setting the parameters found by the regression tree model (SD = standard deviation; N = sample size).

regression tree model has a better predictive capacity than the linear regression model.

In order to find optimal operating conditions, a full factorial design was developed by using the regression tree model as the source of output pressure measurement. The findings showed that linear regression and regression tree models found only four independent variables as significant. It was possible to obtain a significant reduction in the consumption of the steam pressure in line 10, from 9.71 bar defined by the linear regression model to 9.00 bar defined by the regression tree model. The regression tree model demonstrated a clear advantage over the linear regression model because there were better operating conditions and less variability in output pressure.

Because noninvasive experiments such as the assumptions described in this paper reduce the risk of off-quality



12. Paper breaks reported before and after optimization of parameters.

production and downtime, this study demonstrates that the proposed approach of modeling in addition to designed experiments could be implemented for other complex process sections. Further research should be carried out to compare the results found here against results from neural network models. **TJ**

LITERATURE CITED

1. Lawrence, A., Thollander, P., and Karlsson, M., *Sustainability* 10(6): 1851(2018). https://doi.org/10.3390/su10061851.
2. Santos Miranda, M., "The resurgence of the paper industry in the world," *Forbes Mexico*, September 2018. Available [Online] https://www.forbes.com.mx/el-resurgir-de-la-industria-papelera-en-el-mundo/ <03Feb2021>.
3. SFIF, "Global forest industry," The Swedish Forest Industries Federation. Available [Online] https://www.forestindustries.se/forest-industry/statistics/global-forest-industry/.
4. Leslie, F., "Paper manufacturing and products industry," in *Salem Press Encyclopedia*, Salem Press, Hackensack, NJ, 2013.
5. Romero, J.G.I.G., Solís, M.D.V, and Solís, F.M.V., "The business culture and its relationship with learning styles in the pulp, cardboard and paper industry in Mexico," *Investigación Administrativa* 113: 7(2014).
6. Belusso, C., Sawicki, S., Basto Fernandes, V., et al., "A proposal of Infrastructure-as-a-Service providers pricing model using linear regression," *Revista Brasileira de Computação Aplicada* 10(July): (44)2018. https://doi.org/10.5335/rbca.v10i2.7839.
7. Yazir, D. and Sahin, B., *TransNav* (11)4: 635(2017). https://doi.org/10.12716/1001.11.04.09.
8. Herrera, J.B. and Park, Y., *International Journal of Industrial Engineering: Theory, Applications and Practice* 21(4): 179(2013).
9. Muriyatmoko, D. and Putra, L.G.R., *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control* 3(3): 237(2018). https://doi.org/10.22219/kinetik.v3i3.630.
10. Kumari, K. and Yadav, S., *Journal of the Practice of Cardiovascular Sciences* 4(1): 33(2018). https://doi.org/10.4103/jpcs.jpcs_8_18.
11. Abbas, A.T., Alata, M., Ragab, A.E., et al., *Adv. Mater. Sci. Eng.* 2017: 2759020. https://doi.org/10.1155/2017/2759020.
12. Gul, M. and Guneri, A.F., *International Journal of Industrial Engineering: Theory, Applications and Practice* 23(2): 137(2016).
13. Tsai, T.-N. and Tsai, C.-W., *International Journal of Industrial Engineering: Theory, Applications and Practice* 21(1): 19(2014).
14. Adamopoulos, S., Anthony, K., Rapti, E., et al., "Predicting the properties of corrugated base papers using multiple linear regression and artificial neural networks," *Drewno* 59(198): 61(2016).
15. Kilulya, K.F., Mamba, B.B., Ngila, C., et al., *Nord. Pulp Pap. Res. J.* 30(3): 402(2015). https://doi.org/10.3183/npprj-2015-30-03-p402-410.
16. Marklund, A., Hauksson, J.B., Edlund, U., et al., *Nord. Pulp Pap.* Res. J. 13(3): 211(1998). https://doi.org/10.3183/npprj-1998-13-03-p211-219.
17. May-Ostendorp, P.T., Henze, G.P., Rajagopalan, B., et al., *Journal of Building Performance Simulation* 6(3): 199(2013). https://doi.org/10.1080/19401493.2012.665481.
18. Tixier, A.J.-P., Hallowell, M.R., Rajagopalan, B., *Automation in Construction* 69: 102(2016). https://doi.org/10.1016/j.autcon.2016.05.016.
19. Atanasova, N. and Kompare, B., "Modelling of wastewater treatment plant with decision and regression trees," *BESAI 2002—3rd Workshop on Binding Environmental Sciences and Artificial Intelligence* 23(9): 3(2002).
20. Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S., et al., *Landslides* 13: 839(2016). https://doi.org/10.1007/s10346-015-0614-1.
21. Chrysos, G., Dagritzikos, P., Papaefstathiou, I., et al., *ACM Transactions on Architecture and Code Optimization* 9(4): 1(2013). https://doi.org/10.1145/2400682.2400706.
22. Akbari, E., Moradi, R., Afroozeh, A., et al., *Superlattices Microstruct.* 130: 241(2019). https://doi.org/10.1016/j.spmi.2019.04.011.
23. Bosso, M., Vasconcelos, K.L., Ho, L.L., et al., *Journal of Traffic and Transportation Engineering* 7(6): 843(2020). https://doi.org/10.1016/j.jtte.2018.07.004.

24. Choubin, B., Moradi, E., Golshan, M., et al., *Science of the Total Environment* 651: 2087(2019). https://doi.org/10.1016/j.scitotenv.2018.10.064.

25. Jeung, M., Baek, S., Beom, J., et al., *J. Hydrol.* 575: 1099(2019). https://doi.org/10.1016/j.jhydrol.2019.05.079.

26. Shabani, S., Pourghasemi, H.R., and Blaschke, T., *Global Ecology and Conservation* 22: e00974(2020). https://doi.org/10.1016/j.gecco.2020.e00974.

27. Suchetana, B., Rajagopalan, B., and Silverstein, J., *Sci. Total Environ.* 598: 249(2017). https://doi.org/10.1016/j.scitotenv.2017.03.236.

28. Zegler, C.H., Renz, M.J., Brink, G.E., et al., *Agricultural Systems* 180: 102776(2020). https://doi.org/10.1016/j.agsy.2019.102776.

29. Zhang, D., Guo, Y., Rutherford, S., et al., *Sci. Total Environ.* 695: 133758(2019). https://doi.org/10.1016/j.scitotenv.2019.133758.

30. Zhang, L., Traore, S., Ge, J., et al., *Computers and Electronics in Agriculture* 166: 105031(2019). https://doi.org/10.1016/j.compag.2019.105031.

31. Breiman, L., Friedman, J., Olshen, R.A., and Stone, C.J., *Classification and Regression Trees*, Routledge/CRC Press, Boca Raton, FL, USA, 1984. https://doi.org/10.1201/9781315139470.

32. Elith, J., Leathwick, J.R., and Hastie, T., *J. Anim. Ecol.* 77: 802(2008). https://doi.org/10.1111/j.1365-2656.2008.01390.x.

33. Box, G.E.P. and Draper, N.R., *Empirical Model-Building and Response Surfaces*, John Wiley & Sons, New York, 1987.

ABOUT THIS PAPER

Cite this article as:
Rodriguez-Alvarez, J., Lopez-Herrera, R., Villalon-Turrubiates, I., et al., *TAPPI J.* 20(2): 123(2021).
https://doi.org/10.32964/TJ20.2.123

DOI: https://doi.org/10.32964/TJ20.2.123

ISSN: 0734-1415

Publisher: TAPPI Press

About this journal

34. Rodríguez-Álvarez, J.L., "Paper industry," *IEEE Dataport*, 2019. https://dx.doi.org/10.21227/xmma-ky77.

35. Walpole, R.E., Myers, R.H., Myers, S.L., et al., *Probability and Statistics for Engineers*, Pearson Education, London, 1999.

36. Gutierrez-Pulido, H.G. and de la Vara Salazar, R., *Analysis and Design of Experiments*, McGraw-Hill, New York, 2003. [In Spanish]

37. Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L., *Mathematical Statistics with Applications*, Cengage Learning, Boston, 2014.

38. Acikkar, M. and Sivrikaya, O., Turk. *J. Electr. Eng. Comput. Sci.* 26: 2541(2018). https://doi.org/10.3906/elk-1802-50.

ABOUT THE AUTHORS

The topic of this paper corresponds to the use of statistical techniques in complex papermaking sections. The use of predictive models in addition to designed experiments, as described here, is an excellent alternative for solving common problems in papermaking processes. This approach has not been exploited in the rod sizer-spooner section described in the manuscript.

The most difficult aspect of this work was developing a model with high predictive capacity and then using it to find optimal operating conditions. This was solved by using the proposed approach: modeling plus designed experiments. It was interesting to discover that by using the proposed approach, the optimal operating parameters were found without an invasive process.

Mills can use this approach to optimize their processes. The next step would be to use neural network models as an alternative for finding better predictive models.

Rodriguez-Alvarez is a doctoral student; Lopez-Herrera is professor; and Villalon-Turrubiates is professor in the Department of Doctoral Program in Engineering Sciences at Western Technological Institute of Superior Studies in Tlaquepaque, Jal., Mexico. Grijalva-Avila is professor in the Department of Industrial and Manufacturing Engineering, Polytechnic University of Durango, in Durango, Dgo., Mexico. Garcia-Alcaraz is professor in the Department of Industrial and Manufacturing Engineering, Autonomous University of Ciudad Juarez, in Ciudad Juarez, Chih., Mexico. Email Rodriguez-Alvarez at ng718130@iteso.mx.


Rodriguez-Alvarez


Lopez-Herrera


Villalon-Turrubiates


Grijalva-Avila


Garcia-Alcaraz

39. Garcia, C.B., Garcia, J., López Martín, M.M., et al., *J. Appl. Stat.* 42(3): 648(2015). https://doi.org/10.1080/02664763.2014.980789.

40. Hoerl, A.E. and Kennard, R.W., *Technometrics* 12(1): 69(1970). https://doi.org/10.1080/00401706.1970.10488635.

41. Jensen, D.R. and Ramirez, D.E., *International Statistical Review* 76(1): 89(2008). https://doi.org/10.1111/j.1751-5823.2007.00041.x.

42. Lin, F.-J., *Quality & Quantity* 42: 417(2008). https://doi.org/10.1007/s11135-006-9055-1.

43. Piña-Monarrez, M.R., *International Journal of Industrial Engineering: Theory, Applications and Practice* 18(6): (2011).

44. Salmerón Gómez, R., Pérez, J., López Martín, M.D.M., et al., *J. Appl. Stat.* 43(10): 1831(2016). https://doi.org/10.1080/02664763.2015.1120712.

45. de Jongh, P.J., de Jongh, E., Pienaar, M., et al., *ORiON* 31(1): 17(2015). https://doi.org/10.5784/31-1-162.

46. Bera, S. and Mukherjee, I., *Int. J. Adv. Manuf. Technol.* 61: 379(2012). https://doi.org/10.1007/s00170-011-3704-9.

47. Harrington, E.C. Jr., *Industrial Quality Control* 21(10): 494(1965).

48. Ma, X., *Using Classification and Regression Trees: A Practical Primer*, Information Age Publishing, Charlotte, NC, USA, 2018.

49. Duda, R.O., Hart, P.E., and Stork, D., *Pattern Classification*, 2nd edn., John Wiley & Son, New York, 2001.

50. Loh, W.-Y., *WIREs Data Mining Knowl. Discov.* 1(1): 14(2011). https://doi.org/10.1002/widm.8.

51. Antony, J., *Design of Experiments for Engineers and Scientists*, Elsevier, Amsterdam, 2014.