Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

Research Article

Comparative Analysis of Factorial Analysis-Multiple Regression and Random Forest for the Prediction of the Stabilization Time in Furnaces in the Heat Treatment Area in a Metalworking Company

Refugio Chavez Hernandez[1*], Jose Luis Martinez Flores[2], Claudia Malcon Cervera[2], Patricia Cano Olivos[2]

[1] Department of Engineering, University of Monterrey, Nuevo Leon, Mexico
[2] Graduate School in Logistics and Supply Chain Management, Popular Autonomous University of the State of Puebla, Puebla, Mexico
E-mail: refugio.chavez@udem.edu

Received: 10 September 2024; **Revised:** 8 October 2024; **Accepted:** 4 November 2024

Abstract: The purpose of this study is to determine the relationship between the values and attributes of the loads in forged rings made of materials such as nickel, titanium and waspaloy in furnaces for their heat treatment and a prediction model on the preparation and temperature stabilization time of the furnaces, and to be able to carry out the loads and start their holding time according to the recipe assigned for their heat treatment. Applying the Factorial Analysis method, it is possible to identify a reduced number of significant factors that can represent the relationship of the independent variables set, as well as a Multiple Regression Analysis and Random Forest that allows establishing an estimation or prediction system for the time it takes for the furnace to operate the preheating and receive the scheduled load. 6,135 data were collected from full loads in 2020 and six months of 2021, the results had a total variance of 64.61%, a Kaiser-Meyer-Olkin index of 0.620 and a Bartlett sphericity test with a significance of 0.00. The study significantly identifies three important factors on the preparation time of furnaces obtained from the factor analysis, which are: the conditions between loads that represent 33.203% of the total variance, the temperature accuracy for the load 18.149% and the exposure of material in furnaces 13.263%. From the factors identified in the Multiple Regression and Random Forest analysis, it was obtained that the relevant variables are: the temperature difference with respect to the previous load, the weight of the load, the time of holding the load and the treatment temperature for its maintenance have a significant impact on the preparation time. The best prediction method is through the Random Forest algorithm, explaining 95.11% of the variability, its accuracy with respect to the mean square error (MSE) is 6.94 minutes, a mean absolute percentage error (MAPE) of 3.1%, while Multiple Regression manages to explain 77.5% of the variability, a MSE of 69.25 minutes and a MAPE of 9.4%. The result of this research benefits programmers to formulate load sequencing more efficiently in the heat treatment area using the Random Forest algorithm, allowing to increase the productivity and utilization of the furnaces.

***Keywords*:** heat treatment, estimation, factor analysis, multiple regression analysis, random forest

MSC: 62P30, 68T05, 62H25

DOI: https://doi.org/cm.6120255692

1. Introduction

The thermal treatment of the material is one of the fundamental steps so it can achieve the mechanical properties for which it is created [1].

This type of process consists of heating and cooling a metal in its solid state to change its physical-mechanical properties. With the appropriate heat treatment, internal forces can be reduced, grain size, increase toughness or produce a hard surface with a ductile interior [2]. The key for heat treatments consists in the reactions that occur in the material, both in steels and non-ferrous alloys, and occur during the heating and cooling process of the pieces, with established guidelines or times [1].

The high cost of materials, holding time during the process, preparation time between the loads of pieces in furnaces, as well as the imprecision about the sequencing of production scheduling in the heat treatment area, are usually factors highly significant to determine the efficiency of processes, according to Schroeder [3] a company is not competitive if it does not meet high standards in quality, production, times, efficiency, innovation, technology and other relevant concepts that allow organizations stay productive.

For these organizations to be effective and efficient in their processes, their managers must understand and apply some fundamental planning principles for the generation of the product, and also to control the process that originates it [4].

Production scheduling, or sequencing, is an operational response to optimize the production of a good or service. There are several production scheduling techniques [5]. Occasionally, it is complex to define the optimal production scheduling or sequencing technique, this applies to the heat treatment area, first it is necessary to identify the most relevant variables to design and establish a tool or system that allows ensuring the quality of the product treatment and maximize the efficiency of the necessary resources.

The stabilization of a furnace (*the time it takes to reach and maintain a uniform and stable temperature*) depends on several factors. These factors can affect the stabilization time, the accuracy and uniformity of the temperature inside the furnace, some factors are: the furnace design, the energy source, the heat distribution, the thermal load, the quality of the insulation, the temperature control systems, the thermal properties of the materials, the environmental conditions and the frequency of door openings. In this study, the external variables on the furnaces are considered, that is, the conditions and requirements that the load has (*pieces in their heat treatment*), the above generates changes in the parameters to allow the stabilization of the furnace. Basically, the objective is to know and identify the information of the recipe of the charge in the furnace or significant *"variables"* so that the programmers improve the order of the charges reducing the stabilization time in furnaces.

Achieving improved load scheduling allows for multiple benefits, both in terms of operational efficiency and product quality and cost reduction. These benefits are particularly important in industries such as metalworking, the industry where the study is conducted.

2. Methodology

The first step was to review the literature related to the types of heat treatments, followed by Multivariate Analysis (MA) techniques such as the application of factor analysis, multiple regression analysis and random forests.

The second step was to collect the information on the loads in the heat treatment area. For the third step, the data was loaded into the SPSS software to perform the factor reduction and multiple regression test, for the random forests RStudio was used. Finally, the interpretation is presented and discussion of the results, providing information on the decision of the variables selected that will allow the development of an efficient model in the programming of future ring loads in furnaces that reduces the preparation time and minimizes the time the rings stay in the heat treatment area. Figure 1 shows the detailed methodological scheme used in the study.

The computer equipment used for the SPSS Statistics v23 and RStudio 2024.04.2 applications in this work has an Intel Core(TM) i7 8550U @1.80 Ghz CPU and 12 GB of RAM.

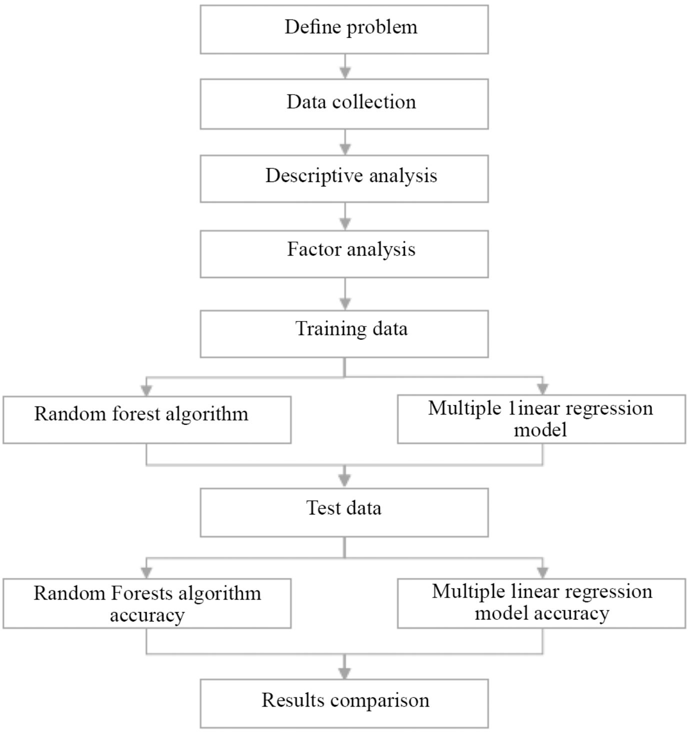


Figure 1. Methodological diagram used

2.1 *Heat treatments*

Heat treatments are processes where only temperature is used as a modifying variable of the microstructure and constitution of metals and alloys, but without changing their chemical composition. The goal of heat treatments is to improve mechanical properties. Thermochemical or surface treatments, additional to use temperature as a variable, modify the chemical composition of a surface layer of the piece due to chemical reactions on the surface of the piece [6].

The heat treatments involve quenching, annealed, tempered, normalized, solubilized, austerized, aged, relieved, stabilized and over-aged.

The difference between each type of treatment consists of fundamental variables such as temperature, time, the way of cooling and the time that the pieces remain in it.

2.2 *Multivariate analysis*

Multivariate Analysis (MA) is a set of statistical and mathematical methods, intended to describe and interpret the data that come from the observation of several statistical variables studied [7].

MA is the part of statistics and data analysis that studies, analyzes, represents and interprets the data that results from observing more than one statistical variable on a sample of individuals. The observable variables are homogeneous and correlated, with none of each predominating. Statistical information in MA is multidimensional in nature, therefore, geometry, matrix calculus and multivariate distributions play a fundamental role [8].

Multivariate information is registered in a data matrix, but often in MA the input information consists of distances matrices or similarities, which measure the degree of discrepancy between individuals.

Multivariate statistical techniques are increasingly used in different branches of science. The exploratory and confirmatory methods, which in most cases are used in combination, require prior knowledge of the studied problem the information available. Although multivariate analysis has its roots in univariate and bivariate statistics, the extension to the multivariate domain introduces additional concepts and questions, ranging from the "theoretical value" to the measurement scales used, the measurement errors, the statistical results of the significance tests and confidence intervals. The use of a multivariate model involves the elaboration of a well-defined research plan that includes the analytical objectives in conceptual terms, the selection of the technique, the evaluation of the basic assumptions of said technique, the estimation of the model and its interpretation, in order to end with the application of validation techniques to determine the stability of the results obtained as the case studies of [9].

[10] in their investigations have classified multivariate methods, but they agree that the three most important aspects to take into consideration with priority order are: the dependence or not between the variables, the measurement scales you use for each of them and the objective pursued in the study.

2.3 *Factor analysis*

The Factor Analysis model is a multiple regression model that relates latent variables with observed variables. Because in many areas it is not possible to directly measure the variables of interest, it is necessary to collect indirect measurements that are related to the concepts of interest. These variables are called latent variables [11].

Factor Analysis has many points in common with principal component analysis, looking for new variables or factors that explain the data. In principal components analysis, there are only orthogonal transformations of the original variables, with emphasis on the variance of the new variables. On the contrary, in factor analysis, it is more interesting to explain the structure of the covariances between the variables [12].

Table 1. Estimation or prediction models and algorithms

Model	Advantages	Disadvantages
Lineal regression	Simplicity and ease of interpretation Efficient computational performance Good performance in linear problems Ease of detecting relationships between variables Multicollinearity control	Assumes a linear relationship Sensitivity to outliers Inability to capture non-linear relationships Unsuitable for categorical data Overfitting in presence of many variables
Decision trees	Easy to understand and interpret Little need for data preprocessing Ability to handle non-linear data Robustness to missing data Fast to train with large data sets	Tendency to overfit Instability Limitations in predictive performance Biased splits Difficulty in managing relationships between variables
Neural networks	Ability to model complex and non-linear relationships Adaptability to different types of data Machine learning of features Resistance to noise in the data	Requires large amounts of data Highly demanding on computational resources Difficult to interpret Overfitting problems Risk of inefficiency if data is not adequate
Time series	Capture temporal dependence Model cyclical and seasonal phenomena Identify trends and structural changes Facilitate the visualization of data over time Ability to analyze recurring events	Requires extensive historical data Susceptibility to unexpected changes Sensitivity to data quality Autocorrelation problems Difficulty in long-term extrapolation
Random forests	Reducing overfitting High accuracy and robustness Handling missing data Ability to handle large data sets and variables	Increased complexity and computational cost Sensitivity to hyperparameter settings Tendency to oversimplify complex relationships Slower predictions

Factor analysis tries to identify underlying variables, or factors, that explain the configuration of correlations within a set of observed variables. Factor analysis is usually used in data reduction to identify a small number of factors that explain most of the variance observed in a larger number of manifest variables [13]. It can also be used to create hypotheses related to causal mechanisms or to inspect variables for subsequent analysis (for example, to identify collinearity before performing a linear regression analysis) [14].

The factor analysis process offers a high degree flexibility, as there are seven factor extraction methods available; and five rotation methods, including direct OBLIMIN and PROMAX for non-orthogonal rotations; three methods available to calculate factor scores; and scores can be saved as variables for further analysis [15].

Once the factor analysis has been carried out and the patterns in the data set have been identified, allowing its dimensionality to be reduced and the factors to be obtained by grouping variables that are highly correlated with each other. The study is based on the exploration of both linear and non-linear methods between the variables, determining the performance with the size of the data set, evaluating the propensity to overfitting without high complexity in the hyperparameters, training speed that allows improving the processing time, fast solution and ease of interpretation, based on the above, the decision is made to check the results on the historical database with the multiple linear regression model and the random forest algorithm. Table 1 shows a comparison between some of the models considered.

2.4 *Multiple linear regression analysis*

Multiple Linear Regression Analysis is a statistical technique used to study the relationship between variables in an extensive variety of situations and to predict diverse phenomena. The simplest regression function is the linear one, where each variable participates in an additive and constant way for all the observed phenomenon [16]. Multiple linear regression brings into the table more than two variables that are explored, and the relationship between the dependent variable and the independent ones is quantified [17].

This allows to establish the relationship that occurs between a dependent variable Y and a set of independent variables ($X1$, $X2$, ... Xk). Multiple linear regression analysis, unlike simple regression, is closer to real analysis situations since social phenomena, events and processes, by definition, are complex and, consequently, must be explained as far as possible by the series of variables that, directly and indirectly, participate in its concretion [18].

When a multiple regression analysis its applied, it is common for both variables, the dependent and independent ones to be continuously measured on the interval or ratio scale. However, this analysis can also be applied when a continuous dependent variable is related to a set of categorical variables; or also in the case that a nominal dependent variable is related to a set of continuous variables [19].

Multiple linear regression techniques start from (k + 1) quantitative variables, with Y being the response variable and ($X1$, $X2$, …, Xk) the explanatory variables [20].

The aim is to extend the techniques of simple linear regression to the 'k' variables. In this line, the variable Y can be expressed through a linear function of the variables ($X1$, $X2$,…, Xk).

$$Y = 0 + 1X1 + 2X2 + \ldots + kXk \tag{1}$$

For this, we will have a probability model (The Normal). The statistician fixes the values of the regressor variables Xki and obtains the corresponding Yi values 'aleatory'.

2.5 *Random Forest*

The Random Forest method is a supervised learning algorithm that is capable of performing regression and classification tasks. The technique is based on grouping and set-based learning, using grouping to concentrate weak sets into a more robust forecasting algorithm through decision trees that are built from several samples. It is equipped to deal with both continuous and qualitative variables [21].

The RF algorithm uses three main hyperparameters that need to be configured before training can begin; node size, amount of trees and amount of sampled features.

Each tree is equipped with:

• A random partition of data (tree bagging)

• A random partition of features (feature sampling)

$$RF = tree\ baggins + feature\ sampling \tag{2}$$

Tree bagging

A B amount of trees can be assembled in the following fashion:

• A B amount of samples are randomly assigned (with replacement) (X, Y) and indexed b (X_b, Y_b) $(b \in \{1..B\})$

• A tree is trained for each pair (X_b, Y_b)

• This reduces the risk for overfitting

Feature sampling

• Each tree can access a limited partition of the total n features, typically $\sqrt{n}$)

• This reduces the correlation between trees, noted by the coefficient ρ

Variance for N trees

The average for B random variables has variance $\dfrac{\sigma^2}{N}$

In the classical hypothesis, where the trees are not independent;

$$V_{forest} = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{3}$$

Split Criteria

Two criteria are used to conduct a branch Split:

• Gini's Criterion

• The Entropy Criterion

Gini's Criterion

• Sum the squares for probability where p and q denote the probability of success and failure, respectively, for each node $(p^2 + q^2)$

• Calculate Gini using the weighted score for each node in the Split

Entropy Criterion

Like in the previous example, p and q are required to be known in each node. The Entropy method can also be used in cases where the target variable is qualitative. The Split with the lowest Entropy is chosen using the father node and other Split variables as a reference. The lower the Entropy, the better [22].

$$Entropy = -p.\log_2(p) - q.\log_2(q) \tag{4}$$

2.6 *Measuring model accuracy*

Having a means to measure the precision of a forecast method is important in allowing the user to observe the error associated to a specific method or algorithm due to uncertainty and variability. When concrete measurements are available, the error is defined as the difference between the forecasted value and the actual measured value.

The Mean Squared Error (MSE) is an objective estimator for the variance associated to term used to describe random errors and given by the following equation:

$$MSE = \frac{SSE}{df_E} = \frac{\sum_{i=1}^{n}(y_1 - \widehat{y}_i)^2}{n-(k+1)} \tag{5}$$

Where y_i denotes observed values, $\widehat{y}_i$ denotes de adjusted values for the dependent variable Y in the i-th iteration. The *MSE* is an average taken over the square error, dividing over the degrees of freedom, and as such can serve as an indicator of how well the regression fits a particular set of data. Taking the square root of the *MSE* can be used as an estimator for the standard deviation or the term associated to random error. Both the *MSE* and it's square root can be used to gauge the magnitude of error in a regression method and can provide insight into the fitting components used [23].

The Mean Absolute Percentage Error *MAPE* is the most useful metric to be used when comparing different methods for their precision, as it compares relative performance [24]. It is commonly used in forecasting methods as a metric for the measurement of precision and is defined by the following equation:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{6}$$

A *MAPE* value under 10% is usually interpreted to be associated with a precise forecasting method, 10-20% indicates a decent amount of precision, 20-50% indicates an acceptable but not completely precise forecast and values exceeding 50% are usually associated with unsuccessful forecasting methods [25].

The R-squared coefficient for multiple regression is similar to it's linear counterpart and is defined as follows:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_1 - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_1 - \bar{y}_i)^2} \tag{7}$$

Where *SST* is the sum of total squares and $\bar{y}$ is the arithmetic mean of the variable Y. The R-squared coefficient measures the percentage of variation in the output variable Y as explained by the variable X. As such, it serves as an indicator for how well a regression method fits a given set of data. The value for the R-square coefficient always exists between 0 and 1, where any value equal or greater than 0.9 being seen as an indicator for a successful regression, with the range of acceptability extending to 0.6. A R-squared value that is lesser or equal than 0.5 is usually taken as an indicator of an unsuccessful regression [26].

3. Case of study

The case of study is applied in FRISA, a company that belongs to the metalworking industry. This company is dedicated to the manufacture of rolled smooth, shaped and open forged products such as bars, hollow bars, discs, rings, rollers, blocks and screeds for customers worldwide. The study takes place on the Aerospace spacecraft. The materials used to manufacture the rings are super-alloys of titanium, nickel and waspaloy. The growth in the production of rolled rings for the aerospace industry with more than 1,000 Sku's, has grown exponentially, to meet their demand, the company has emphasized on being more efficient in their processes.

In this particular case, the aim is to determine the vital variables that affect the preparation and stabilization time of furnaces in the heat treatment area. Each ring belongs to a standard or model, which requires a recipe for heat treatment that meets the customer's specifications. The first process is the forming of the ring and begins with a process of cutting a specific alloy piece, then the piece is sent to an oven for its necessary soaking, in the next process, the piece is sent to a press and with a punch the perforation is made giving the shape of a ring.

Next is the rolling process according to the ring standard, requiring a sequence of "*n*" rolling cycles depending on the standard, it begins with the ring in the oven until it reaches its cycle time, once the heating is complete, it is sent to the rolling equipment, there the rings are worked in three dimensions, height, internal and external diameter, the process is repeated until the standard cycles are completed.

The post-forging process is the one that is done in the area of interest, the heat treatment, just as the forging process has "*n*" sequences to complete its recipe, in each of the sequences the type of heat treatment required by the standard is assigned. In Figure 2, the Layout of the heat treatment area is shown, which has 4 furnaces, 1 water cooling tub, 1 polymer cooling tub, 2 forced air chambers and cooling to room temperature can be performed anywhere in the ship.

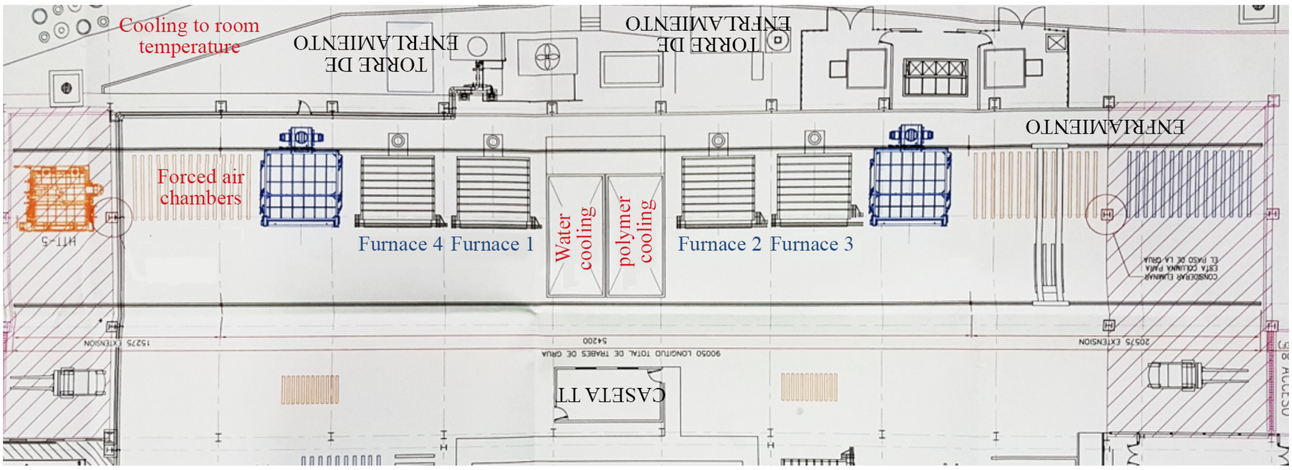


Figure 2. Heat treatment area layout

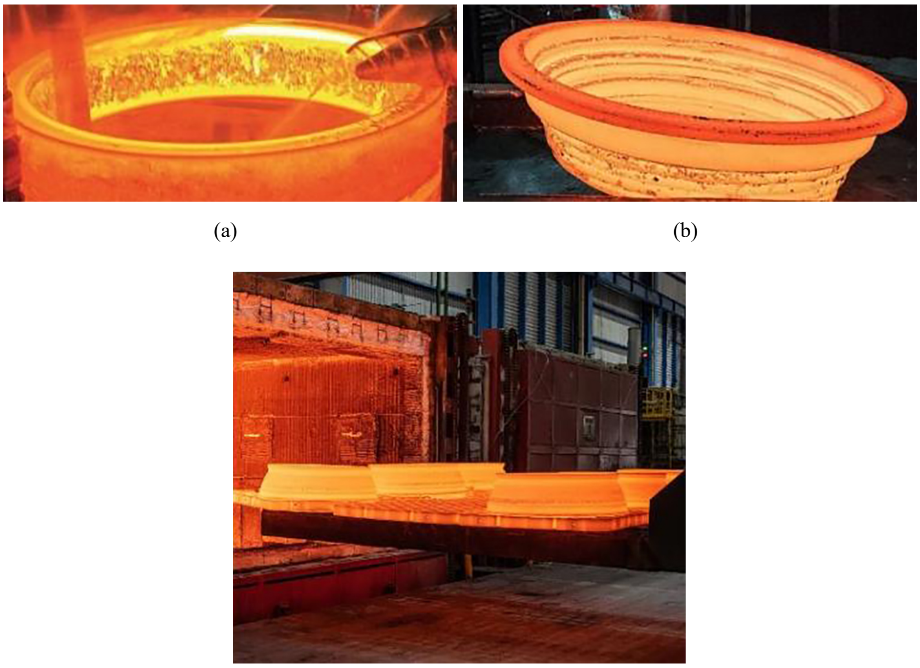
It has been possible to determine that the variables that intervene or affect the time of preparation or preheating time in ovens are: the type of the loads heat treatment, the number of pieces or batch size, the weight of the load, the hardness of the alloy, the treatment temperature for its support, the tolerance of the furnace for its stabilization, the time of the load and the difference in temperature with respect to the previous load that was in the furnace.

Table 2. Types of heat treatment and their classification

Type of treatment	Minimum temperature (°C)	Maximum temperature (°C)	Temperature range (°C)	Classification
Solubilized	885	1,200	315	1
Temple	875	1,149	274	2
Stabilization	782	850	68	3
Over-Aged	676	704	28	4
Austenitized	650	1,170	520	5
Normalization	600	1,060	460	6
Annealing	600	1,200	600	7
Aging	552	1,080	528	8
Tempering	400	790	390	9
Relieving	400	1,177	777	10

Table 2, shows the types of heat treatment that can be assigned to each standard in each of the sequences required to meet customer specifications, each type of treatment requires a temperature and this might vary.

Our first variable is the type of treatment in sequence "*n*", evaluated by its classification, which is determined from the highest temperature of the minimum temperatures and the lowest temperature range of the type of treatment.



(a) (b)

(c)

Figure 3. (a) Straight ring, (b) Shaped ring and (c) Loading rings into furnace

The second variable is the batch size or number of pieces in the load, the lot is limited by the furnace capacity due to the diameter of the rings, in the case where the number of pieces exceeds the furnace capacity the load is divided, Figure 3 shows a straight ring standard, a shaped ring standard, and the loading of a batch of shaped rings for their assigned furnace heat treatment.

Table 3. Ring alloys hardness and its category

Material	Hardness	Category	Material	Hardness	Category	Material	Hardness	Category
Alloy 718	410	1	304	325	13	Alloy C-263	248	25
Thermo-Span	400	2	Ti 6-2-4-2	318	14	Alloy 230	241	26
Alloy 901	388	3	Alloy PE16	318	15	H-46	240	27
17-4 PH	363	4	M-152	311	16	Ti 834	225	28
15-5 PH	363	5	Waspaloy	302	17	321	207	29
Rene 41	363	6	Alloy 80 A	270	18	Alloy 625	190	30
Ti 21 s	360	7	Alloy 909	269	19	Alloy 617	173	31
Alloy 188	341	8	Hykro	269	20	Alloy 242	158	32
Alloy 718 Plus	340	9	CMV	256	21	418	125	33
PH 15-7 Mo	335	10	347	255	22	Alloy 907	102	34
Ti 6-4	334	11	FV-535	250	23	Alloy X	89	35
Ti 6-4 Pre	334	12	Alloy 282	250	24	Alloy 718 Pre	36	36

The third variable is the weight of the load, once the batch has been divided if necessary, the total weight of the load is obtained expressed in tons.

The fourth variable is the hardness of the alloy, which is measured through the Brinell method. Table 3 shows the hardness of each alloy that is treated.

The fifth variable of interest is the temperature in Celsius degrees of the recipe for the treatment of the standard, the oven where the load will be carried out must be within parameters to begin its holding time, Table 4 shows the ranges by category, this allows you to quickly identify the oven that can handle the load.

Table 4. Category assigned according to its temperature range

Category	Range	
1	400	500
2	500	600
3	600	700
4	700	800
5	800	900
6	900	1,000
7	1,000	1,100
8	1,100	1,200

The sixth variable is the temperature tolerance in the oven for its stabilization, also in units of Celsius degrees, the ovens have a capacity to set the temperature in a range, which depends on the data expressed in the recipe. The company has four furnaces to subject the rings to their heat treatment, the furnaces have the same volumetric capacity, but have a different precision in their tolerance when stabilizing the furnace, furnace number 4 has a lower tolerance than makes it more exact than the rest of the ovens, Table 5 shows the tolerance with which each oven has.

Table 5. Temperature tolerance capacity in furnaces and its category

Furnace	Furnace (°C)	Category
Furnace TT 4	± 5	1
Furnace TT 3	± 10	2
Furnace TT 1	± 15	3
Furnace TT 2	± 20	4

The seventh variable is the steeping time or holding time of the rings in the oven, the assigned recipe contains the time that will be exposed to the conditions necessary for its treatment, the unit of the stay time is in hours. The eighth variable is the value of the difference in Celsius degrees between loads assigned to each of the ovens, depending on whether the need is to increase or reduce the temperature in the oven to stabilize it for the next load.

Next, in Table 6 an example of a batch of three pieces of standard 219 is described for its heat treatment, it shows the detail of the information of the recipe 57,538 for its first sequence requires an annealing.

Table 6. Description of sequence 1 for a 219 standard

Sequence	Standard	Recipe	Material	Hardness	Type of Treatment	Amount of pieces	Tons for batch
1	219	57,538	Rene 41	335	Annealing	3	0.312

In Table 7, the necessary conditions for the annealing sequence are detailed, it is assigned to furnace number three, with two hours of holding at 1,079 degrees with a tolerance of ± 10 Celsius degrees.

Table 7. Sequence 1 condition for a standard 219 in oven

Furnace	Time of holding [hrs]	Temperature °C	Tolerance temperature °C
Furnace TT 3	2	1,079	± 10

After the time of holding or soaking the rings in their first sequence, it passes to its cooling medium, which, for our study, is not considered relevant since it is a subsequent and independent process for the preparation of the oven in its stabilization time. The information is displayed in the Table 8 just for reference to show the complete cycle of the sequence.

Table 8. Sequence 1 condition for a 219 standard in its cooling medium

Cooling medium	Time of cooling [hrs]
Forced air	0.55

Once the first loading sequence is finished, it returns to the beginning of the cycle for the second treatment depending on the loads already programmed, it is considered a first inputs, first outputs (FIFO) system, sometimes they are considered emergencies depending on the time that the pieces are retained by various factors, such as quality review, assignment of a new recipe by the metallurgy area, parts delayed by processes prior to heat treatment or priority by the customer. Tables 9, 10 and 11 show the information for the second sequence of standard 219.

Table 9. Description of sequence 2 for a 219 standard

Sequence	Standard	Recipe	Material	Hardness	Type of treatment	Amount of pieces	Tons for batch
2	219	58,028	Rene 41	335	Solubilized	3	0.312

Table 10. Sequence 2 condition for a standard 219 in oven

Furnace	Soak time [hrs]	Temperature °C	Tolerance temperature °C
Furnace TT 1	0.5	1,066	± 15

Table 11. Sequence 2 condition for a 219 standard in its cooling medium

Cooling Medium	Time of cooling [hrs]
Forced air	0.68

Tables 12, 13 and 14 show the information for the third sequence of standard 219.

Table 12. Description of sequence 3 for a 219 standard

Sequence	Standard	Recipe	Material	Hardness	Type of treatment	Amount of pieces	Tons for batch
3	219	58,033	Rene 41	335	Aging	3	0.312

Table 13. Sequence 3 condition for a standard 219 in oven

Furnace	Soak time [hrs]	Temperature °C	Tolerance temperature °C
Furnace TT 2	14.48	760	± 20

Table 14. Sequence 3 condition for a 219 standard in its cooling medium

Cooling medium	Time of cooling [hrs]
Air	5.45

First, a descriptive analysis of the 6,135 loads collected information in the heat treatment area in 2020 and six months of 2021, only for eight variables mentioned above, having as study variable the time of preparation of the ovens for the load. In Table 15, the descriptive statistics of the preheating time in ovens are shown, regarding the average time in the preparation of ovens for the load to be carried is 69.87 minutes and standard deviation of 13.21 minutes, there is a range between the minimum time and maximum of 100.41 minutes, which allows us to consider that there is an area of opportunity to reduce the stabilization time.

Table 15. Descriptive statistics of ring loads in heat treatment

Time of preparation [min]	
Media	69.8762476
Typical error	0.16875132
Median	71.3375976
Standard deviation	13.2176569
Sample variance	174.706454
Kurtosis	1.4352765
Asymmetry coefficient	0.7514041
Range	100.418611
Minimum	47.5706174
Maximum	147.989228
Sum	428,690.779
Count	6,135

In Figure 4, the histogram of the preparation time behavior is shown, it can be noted that the greatest number of the ovens preparation are between the range of 53 to 83 minutes, very few frequencies with times above 90 minutes. For the improvement opportunity, it is observed that there are conditions that allow the oven to prepare in a time between 48 to 55 minutes, our study aims to determine the variables that allow to explain the variation in the preparation time, and generate a prediction model based on our significant variables that predict the stabilization time and reorder or sequence the loads in a way that reduces the time used in the furnaces preparation.

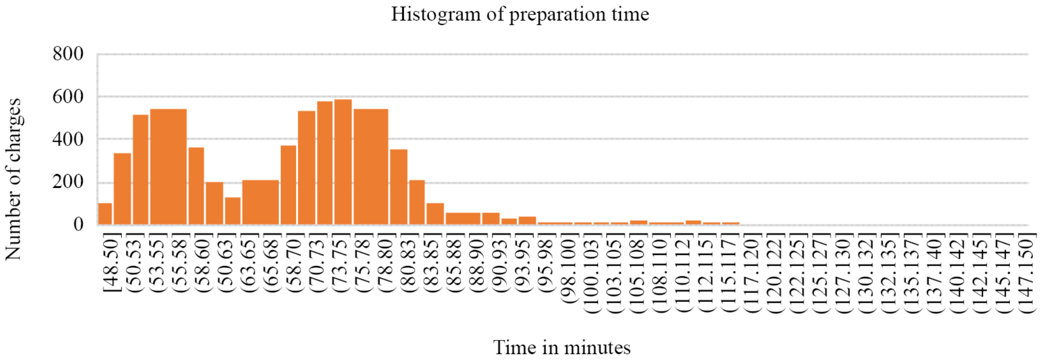


Figure 4. Histogram of the stabilization time in ovens

Table 16 shows the number of loads served by each of the furnaces in the heat treatment area, as well as its average time in the furnace preparation, it is observed that furnace number 2 has the lowest average preparation time with a high number of loads, unlike the oven number 1, which has the highest average time and the least number of loads.

Table 16. Number of loads and average prep time per furnace

Furnace	Loads	Average time
Furnace TT 1	1,169	73.5395706
Furnace TT 2	1,878	64.2444887
Furnace TT 3	1,200	71.6849499
Furnace TT 4	1,888	72.0603449

3.1 *Results of the factor analysis*

Once the descriptive analysis has been applied, the factor analysis of principal components was applied using the VARIMAX rotation method, an orthogonal rotation method that allows minimizing the number of variables that have high loads on each factor and simplifies the factors interpretation in our variables involved in the heat treatment sequences, through this method we want to identify a relatively small number of factors that can be used to represent the relationship that exists in our set of inter-correlated variables. The SPSS V23 package was used for the analysis.

KMO and Bartlett test		
Kaiser-Meyer-Olkin measure of sampling adequacy		0.620
	Chi-square approx	16,665.519
Bartlett's test of sphericity	df	28
	Significance	0.000

Figure 5. SPSS, KMO and Bartlett test results

Figure 5 shows the value of the Kaiser-Meyer-Olkin (KMO) index obtained from the test with a value of 0.620. The KMO index varies between 0 and 1, when the value of the indicator is greater than 0.9, it can be considered as an excellent value for factor analysis, however, if the value is below 0.5, it can be interpreted as not suitable for analysis, indicating that the data may not be appropriate for performing a factor analysis, since the correlations between the variables are not strong enough to group them. In our case, the KMO value is 0.620, considering it as acceptable, it can be stated that the correlations between pairs of variables are explained by other variables, therefore, the factor analysis is appropriate. In the Bartlett's test of sphericity, given the critical value (significance) is less than 0.05, we can affirm that the correlation matrix is not an identity matrix.

After applying the VARIMAX rotation method, three important and significant factors were found with an auto value greater than one and represented 64.615% of the total variance for subsequent analysis, as shown in Figure 6.

Total variance explained			
Component	Initial eigenvalues		
	Total	% of variance	% cumulativ
1	2.656	33.203	33.203
2	1.452	18.149	51.352
3	1.061	13.263	64.615
4	0.947	11.833	76.449
5	0.741	9.259	85.708
6	0.659	8.243	93.951
7	0.396	4.948	98.899
8	0.088	1.101	100.00
Extraction method: principal component analysis			

Figure 6. SPSS results, total variation

Figure 7 represents the rotated component matrix that shows that three factors were extracted that have a factor load value greater than 0.5. The factor with less than three variables is generally weak and unstable.

In addition, the factor loading of the variables must be greater than 0.5 to determine which elements will be grouped into which factors. If the value is 0.5, it depends on the highest factor load assigned by each of them. By standards, the value must be greater than 0.5. Therefore, it can be concluded that the data are suitable for further analysis.

Rotated component matrix[a]			
	Component		
	1	2	3
Treatment temperature	0.938	-0.075	-0.030
Type of treatment	-0.927	-0.119	0.091
Difference of temperature between loads	0.809	-0.068	0.059
Alloy hardness	0.005	0.780	-0.032
Furnace tolerance	0.032	-0.719	0.043
Number of pieces	0.074	-0.193	0.684
Tons of loads	0.007	0.493	0.600
Soak time	-0.458	0.024	0.534
Extraction method: principal component analysis			
Rotation method: Varimax with Kaiser normalization			
a. The rotation has converged in 4 iterations			

Figure 7. SPSS results, rotated component matrix

The result showed that of the eight variables they were reduced to three factors, these three factors represent 64.615% of the total variance. The three factors are described as (Table 17):

Table 17. The total variance of three factors

Factor I: Conditions between loads	Factor II: Temperature precision for the load	Factor III: Matter exposure
This factor consists of three attributes, the temperature and the treatment type of the next programmed load in the furnace and configuration of the previous load. They represent 33.203% of the total variance.	This factor consists of only two attributes, the alloy hardness and the tolerance on furnace temperature for its stabilization. They represent 18.149% of the total variance.	This factor consists of three attributes, the batch size to be treated, the load weight and holding time in the furnace. They represent 13.263% of the total variance.

3.2 *Results of multiple regression analysis*

The intention of this test is to consider more than one explanatory variable, to understand the functional relationship between our dependent variable, the preparation or preheating time, and the independent variables, the type of the loads heat treatment, the number of pieces or size of batch, the load weight, the alloy hardness, the treatment temperature for its support, the furnace tolerance for its stabilization, the load time and the temperature difference based on the previous load that it was attended to in the oven, and to study what are the causes of the variation in the ovens preparation time. A summary of the multiple regression results is presented in Figures 8, 9 and 10.

Model Summary				
Model	R	R-squared	Adjusted R-squared	Standard error of the estimate
1	0.755[a]	0.600	0.600	8.36423

a. Predictors: (Constant), Type of treatment, Tons of loads, Number of pieces, Furnace tolerance, Alloy hardness, Soak time, Difference of temperature between loads, Treatment temperature

Figure 8. SPSS results, model summary

ANOVA[a]					
Model	Sum of squares	df	Mean square	F	Sig.
Regression	643,072.671	8	80,384.084	1,148.995	0.000[b]
Residual	428,576.957	6,126	69.960		
Total	1,071,649.628	6,134			

a. Dependent variable: Furnace preparation time
b. Predictors: (Constant), Type of treatment, Tons of loads, Number of pieces, Furnace tolerance, Alloy hardness, Soak time, Difference of temperature between loads, Treatment temperature

Figure 9. SPSS results, ANOVA

Coefficients[a]					
Model	Unstandardized coefficients		Standardized coefficients	t	Sig.
	B	Standard error	Beta		
(Constant)	55.718	1.929		28.89	0.000
Alloy hardness	0.01	0.002	0.056	6.119	0.000
Soak time	-0.652	0.024	-0.251	-26.969	0.000
Tons of loads	6.87	0.128	0.458	53.556	0.000
Treatment temperature	0.012	0.002	0.148	7.346	0.000
Furnace tolerance	-0.209	0.018	-0.096	-11.347	0.000
Difference of temperature between loads	-0.042	0.001	-0.714	-64.587	0.000
Number of pieces	-0.001	0.022	-0.001	-0.065	0.948
Type of treatment	-0.269	0.078	-0.068	-3.443	0.001

a. Dependent variable: Furnace preparation time

Figure 10. SPSS results, coefficients

Figure 9 shows the results and indicate that there is a strong and significant relationship between the preparation time in ovens and the independent variables previously exposed (F = 1,148.995, with a significance < 0.05). Figure 8 shows the value $R2$ = 0.600, indicating that the independent variables explained 60 percent of the variation in the preparation or preheating time in ovens.

Figure 10 shows the individual determinants, it was found that the temperature difference between loads (β = -0.714, t = -64.587, p = 0.000 < 0.05). This was followed by the weight of the load (β = 0.458, t = 53.556, p = 0.000 < 0.05), the holding time or soak (β = -0.251, t = -26.969, p = 0.000 < 0.05) and the treatment temperature (β = 0.148, t = 7.346, p = 0.000 < 0.05). All the variables are found with a significance lower than 0.05, except for the number of pieces' variable, so the multiple regression was performed again, eliminating the variable in the new test.

Coefficients[a]

Model	Unstandardized coefficients		Standardized coefficients	t	Sig.
	B	Standard error	Beta		
(Constant)	55.711	1.925		28.940	0.000
Alloy hardness	0.010	0.002	0.056	6.126	0.000
Soak time	-0.652	0.024	-0.251	-26.981	0.000
Tons of loads	6.870	0.128	0.458	53.615	0.000
Treatment temperature	0.012	0.002	0.148	7.346	0.000
Furnace tolerance	-0.209	0.018	-0.096	-11.349	0.000
Difference of temperature between loads	-0.042	0.001	-0.714	-64.603	0.000
Type of treatment	-0.269	0.078	-0.068	-3.443	0.001

a. Dependent variable: Furnace preparation time

Figure 11. SPSS results, coefficients

ANOVA[a]

Model	Sum of squares	df	Mean square	F	Sig.
Regression	643,072.371	7	91,867.482	1,313.350	0.000[b]
Residual	428,577.256	6,127	69.949		
Total	1,071,649.628	6,134			

a. Dependent variable: Furnace preparation time
b. Predictors: (Constant), Type of treatment, Tons of loads, Furnace tolerance, Alloy hardness, Soak time, Difference of temperature between loads, Treatment temperature

Figure 12. SPSS results, ANOVA

Model Summary				
Model	R	R-squared	Adjusted R-squared	Standard error of the estima
1	0.755[a]	0.600	0.600	8.36355

a. Predictors: (Constant), Type of treatment, Tons of loads, Furnace tolerance, Alloy hardness, Soak time, Difference of temperature between loads, Treatment temperature

Figure 13. SPSS results, model summary

According to the setting of Figure 11, the following preparation time prediction model can be expressed:

Furnace preparation time = 55.711-0.010 alloy hardenss -0.652 soak time + 6.870 tons of loads + 0.012 treatment temperature -0.209 furnace tolerance -0.042 difference of temperature between loads -0.269 type of treatment.

Figure 12 shows the results and indicate that there is a strong and significant relationship between the preparation time in ovens and the independent variables previously exposed (F = 1,313.350, with a significance < 0.05). Figure 13 shows the value $R2$ = 0.600, indicating that the independent variables explained 60 percent of the variation in the preparation or preheating time in ovens. Figure 11 shows the individual determinants, it was found that the temperature difference between loads (β = -0.714, t = -64.603, p = 0.000 < 0.05). This was followed by the weight of the load (β = 0.458, t = 53.615, p = 0.000 < 0.05), the time of holding or soaking (β = -0.251, t = -26.981, p = 0.000 < 0.05) and the treatment temperature (β = 0.148, t = 7.346, p = 0.000 < 0.05). All the variables are found with a significance lower than 0.05.

3.3 *Precision test for predicted preparation times using a linear regression*

In order to carry out an evaluation through the linear regression method, data from the 6,135 samples was grouped into training and test sets. Approximately 70% of the data was assigned to the training samples (4,294 data entries), 30% was assigned to the test set. Historical loads for thermal treatment were verified in 1,841 instances. The model's performance was measured through the Mean Square Error and Mean Absolute Percentage Error methods.

The mean square error associated to the oven preparation times as per the linear regression method is 69.25 minutes per load. The associated standard deviation is 8.32 minutes and the estimator used to measure the model's performance gives a 9.4% overall error in the oven preparation times. A tool like the linear regression model is a critical instrument in the field of thermal treatment as it allows an opportunity for visualization of data in the decision making process of defining load sequences, incorporating several variables into an integral vision of an optimized oven usage.

3.4 *Random forest analysis*

The data set for historical oven loads in the thermal treatment sector has normalized variables as does the linear regression model. 4,294 data samples stored in the train_data variable in RStudio were selected at random and used for both previously mentioned applications. The remaining 1,841 data samples were stored in the test_data variable, and will in turn be used to measure the algorithm's performance. Figure 14 displays the stored matrices associated to variables data, modelo, rf_model, test_data and train_data which are used in RStudio for data set treatment.

R	Global Environment
Data	
data	6,135 obs. of 8 variables
modelo	Large 1m (12 elements, 2.1 MB)
rf_mode1	List of 18
test_data	1,841 obs. of 8 variables
train_data	4,294 obs. of 8 variabl es

Figure 14. Environment variables in RStudio

Incorporating data about the main variables allows the user to execute a Random Forest algorithm in the RStudio platform. The RF model is executed on the training data set that contains information related to the parameters that are predetermined in the RStudio environment. Figure 15 describes the result of executing the training model of the Random Forest regression algorithm, as shown in the indexed Type of Random Forest: Regression (case $N = 100$), number of variables tried at each Split = 2. The algorithm considers two random variables to determine the best possible classification, allowing the Random Forest algorithm to be modeled with a high diversity of trees. Mean of Squared Residuals = 8.523845 is the measured MSE associated to the residues, where the residues themselves indicate the average over the squared error of the values estimated by the algorithms and the real values. % Var explained = 95.11 indicates the percentage of variability for the data sets the model incorporates into it's analysis and therefore shows that the chosen model can account for 95.11% of the variability in the furnace preparation time, a solid fitting.

Call:	
randomForest (formula = Time_prep_furnace ~ . , data = train_data, ntree = 100)	
Type of random forest:	regression
Number of trees:	100
No. of variables tried at each split:	2
Mean of squared residuals:	8.523845
% Var explained:	95.11

Figure 15. Results of training using the Random Forest algorithm in the RStudio plaform. Dataset extracted from the historial load records

Figure 16 shows the relevance of the predicting variables, which outline the most important variables in determining the oven preparation times for subsequent loads, according to the Random Forest algorithms, where *Diff_temp_between_loads* is observed to indicate the amount by which a given variable (temperature) will need to be increased/decreased. The parameter used to measure the relevant of independent variables in the model forecast is IncNodePurity (INP), which measures the "purity" of each node in decision trees, and refers to the minimization of MSE in the case of regression models. The higher the value for INP, the more relevant the variable for the overall model.

	IncNodePurity
Alloy hardness	32,698.87
Soak time	85,804.85
Tons of loads	164,920.39
Treatment temperature	61,035.26
Furnace tolerance	21,634.59
Difference of temperature	331,864.82
Type of treatment	19,424.69

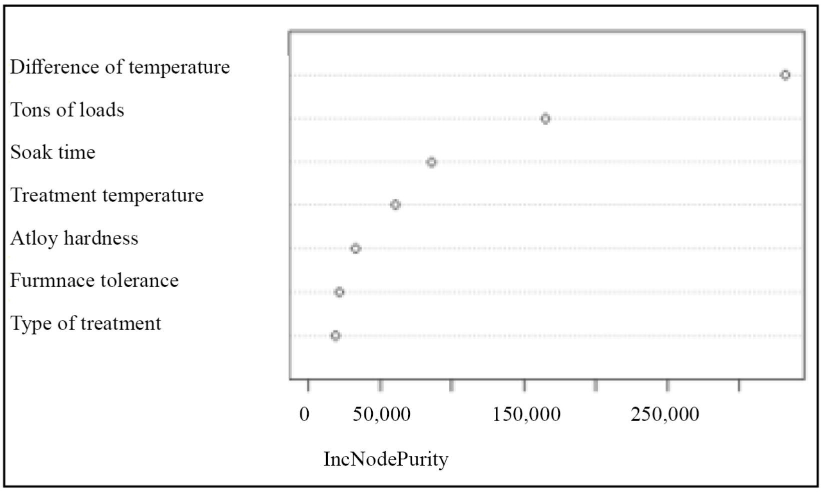


Figure 16. Parameter to determine predicting variables in the Random Forest model. RStudio platform

The MSE for the oven preparation times according to the Random Forest algorithm is 6.94 minutes per load, the standard deviation is approximately 2.633 minutes of furnace preparation time, and the model's performance is measured at 96.9% accuracy.

Table 18 shows a comparison of results through both methods under consideration in this study.

Table 18. Comparison of results on test data from Multiple Regression and Random Forest

	Multiple linear regression	Random forest
Variable Relevance	1. Temperature difference between loads 2. Tons of loads 3. Soak time 4. Treatment temperature 5. Furnace tolerance 6. Treatment type 7. Alloy hardness	1. Temperature difference between loads 2. Tons of loads 3. Soak time 4. Treatment temperature 5. Alloy hardness 6. Furnace tolerance 7. Treatment type
% Accounted Variability	77.5%	95.11%
Test Precision	MSE = 69.25 MAP E = 9.4%	MSE = 6.94 MAPE = 3.1%
Computational burden	13.6% CPU used 226.7 MB RAM 3 seconds to solve	15.9% CPU used 405.3 MB RAM 6 seconds to solve

4. Closing remarks

Three relevant factors have been identified throughout this study; precision in the measurement for temperature between loads, load temperature and exposition of dough in furnace. Furthermore, a correlation has been established between temperature difference between loads, tons of loads (*weight*), soak time and treatment temperature for both methods. There are some limitations to this study. Only 8 independent variables are considered: Treatment Type, Amount of Items, Tons Load, Allow Hardness, Treatment Temperature, Furnace Tolerance, Soak Time and Difference of Temperature Between Loads. Some factors are excluded on account of not being relevant to the present analysis. Furnace preparation time, or the time to pre-heat a furnace, is the dependent variable that makes up this study through either method. The studied phenomenon can be further studied in future works through the incorporation of additional dependent variables that can offer further insight. This study is limited to the Thermal Treatment department of the FRISA Aerospace facilities. A similar study can be executed in other smelting facilities. The results presented in this study are relevant to determining the importance of the independent variables in thermal treatment procedures with an universal goal of dismissing process time and associated costs. Both methods used, Random Forest and Multiple Regression, agree in the relevance of the predicting variables for the model, with the Random Forest method offering a more exact estimation of the oven preparation times through the reduced risk of overfitting.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Gave Chagua JL. Heat treatment of 1050a aluminum strip from continuous casting and hot rolling. *Revista Universidad y Sociedad*. 2022: 610-616.

[2] Ma Y, Yang J, Liu Z, Rong C, Zhou L, Liu Y. Directional heat treatment technique to prepare columnar grains in a TiAl alloy: Microstructure evolution and deformation behavior. *Journal of Materials Processing Technology*. 2024; 330: 118462. Available from: https://doi.org/10.1016/j.jmatprotec.2024.118462.

[3] Schroeder RG. *Operations Research*. 1st ed. España: Mcgraw-hill Interamericana de España S.L; 2011.

[4] Chapman SN. *Planeación y Control de la Producción*. 1st ed. México: Pearson Education; 2006.
[5] Wang S, Zhang X. Production scheduling of prefabricated components considering delivery methods. *Scientific Reports*. 2023; 13: 15094. Available from: https://doi.org/10.1038/s41598-023-42374-w.
[6] Huircapan MP, Franco M, Rosa HD. Effect of microstructure and heat treatment of hypereutectic Al-Si alloys obtained by Rapid Solidification. *Revista Matéria (Rio de Janeiro)*. 2022; 27(02). Available from: https://doi.org/10.1590/1517-7076-RMAT-2021-49007.
[7] Pérez C. *Análisis Multivariante de Datos [Multivariate Data Analysis]*. 1st ed. Madrid, España: Garceta grupo editorial; 2013.
[8] Cuadras C. *Nuevos métodos de análisis multivariante [New Methods of Multivariate Analysis]*. 1st ed. Barcelona: CMC Editions; 2019.
[9] Botero-Acosta A, Amanda L, Sagan V, Demir I, Muste M, Paul B, et al. Multivariate analysis and anomaly detection of a us reservoir sedimentation data set. *Journal of Hydrologic Engineering*. 2024; 29(5): 04024031. Available from: https://doi.org/10.1061/JHYEFF.HEENG-6206.
[10] Addis R, Mannu A, Pintore G, Petretto GL. Statistical analysis applied to the production of mirto liqueur. *Applied Sciences*. 2024; 14(14): 5973. Available from: https://doi.org/10.3390/app14145973.
[11] Peng YZ, Li SY, Hu SC, Wu ZK, Tang WZ, Wang ZJ. Analysis of factors influencing construction and operation of Ningxia's Internet + urban and rural water supply project and management recommendations. *Journal of Tsinghua University*. 2024; 14(14): 243-249. Available from: https://doi.org/10.16511/j.cnki.qhdxxb.2023.22.051.
[12] Morales P. *El Análisis Factorial en la construcción e interpretación de tests, escalas y cuestionarios [Factor Analysis in the Construction and Interpretation of Tests, Scales, and Questionnaires]*. Madrid: Universidad Pontificia Comillas; 2011.
[13] Ma H, Yang X, Qi K. Research on the enhancement path of talent cultivation in the integration of dual-creation education and professional education for teacher training majors in colleges and universities based on multivariate statistical analysis. *Applied Mathematics and Nonlinear Sciences*. 2024; 9(1). Available from: https://doi.org/10.2478/amns.2023.2.01315.
[14] Chang ZK, Chen YW, Feng X. Multi factor analysis method of traffic accidents based on driving behavior. *Advances in Transportation Studies*. 2022; 1: 57-66. Available from: https://doi.org/10.53136/97912599481136.
[15] Corporation IBM. *Análisis factorial explicativo*. USA: SPSS Statistics; 2021.
[16] Manala K, Nheta W. Application of multiple linear regression (MLR) analysis on the concentration of chromite plant tailings by a shaking table. In: *Proceedings of the World Congress on Mechanical, Chemical, and Material Engineering*. Barcelona, Spain; 2024. p.1–8. Available from: https://doi.org/10.11159/mmme24.117.
[17] Kar S, Pattnaik S, Sutar MK. Regression analysis and morphological study of chemically treated organic fibers. In: Sahoo S, Yedla N. (eds.) *Recent Advances in Mechanical Engineering*. Singapore; 2024. p.279–291.
[18] Laqua A, Schnee J, Pletinckx J, Meywerk M. Identification of factors influencing the riding experience on e-mountain bikes: An analysis of the rider-bicycle interaction. *Transportation Research Part F: Traffic Psychology and Behaviour*. 2023; 98: 61-72. Available from: https://doi.org/10.1016/j.trf.2023.08.008.
[19] Santana A, Rama J. *Manual de Análisis de Datos con Stata [Data Analysis Handbook using Stata]*. 1st ed. Madrid: Tecnos; 2017.
[20] Aragoncillo AM, Cleary D, Thayasivam U, Lomboy G. Water sorptivity prediction model for concrete with all coarse recycled concrete aggregates. *Construction and Building Materials*. 2023; 394: 132128. Available from: https://doi.org/10.1016/j.conbuildmat.2023.132128.
[21] Biernat E, Lutz M. *Data Science: Fundamentals and Case Studies: Machine Learning with Python and R*. 1st ed. Saint-Germain: Eyrolles; 2015.
[22] Hemal SH, Khan MAR, Ahammad I, Rahman M, Khan M, Ejaz S. Predicting the impact of internet usage on students' academic performance using machine learning techniques in Bangladesh perspective. *Social Network Analysis and Mining*. 2024; 14(1): 66. Available from: https://doi.org/10.1007/s13278-024-01234-9.
[23] Wang ZD, Meng YL, Li Y, Gao H, Zhang T, Pan GM, et al. Inverse design of ultranarrow and high-efficiency color filters based on tandem convolutional neural networks. *Optics Communications*. 2024; 573: 130995. Available from: https://doi.org/10.1016/j.optcom.2024.130995.
[24] Makridakis SG, Wheelwright SC, Hyndman RJ. *Forecasting Methods and Applications*. 3rd ed. New York: Wiley; 1997.

[25] Lewis CD. *Industrial and Business Forecasting Methods*. London: Butterworths; 1982.
[26] Ostertagová E, Ostertag O. Regression analysis and seasonal adjustment of time series. *Journal of Automation and Control*. 2015; 3: 118-121.