


MULTIPLE RESPONSE OPTIMIZATION: COMPARATIVE ANALYSIS BETWEEN MODELS OBTAINED BY ORDINARY LEAST METHOD AND GENETIC PROGRAMMING

Nilo Antonio de Souza Sampaio^A, José Salvador da Motta Reis^B, José Glenio Medeiros de Barros^C, Cleginaldo Pereira de Carvalho^D, Fabricio Maciel Gomes^E, Luís César Ferreira Motta Barbosa^F, Messias Borges Silva^G



| ARTICLE INFO | ABSTRACT |
|--|---|
| <p>Article history:</p> <p>Received 08 May 2023</p> <p>Accepted 03 August 2023</p> | <p>Purpose: This work aims to analyze and compare the performance between the Ordinary Least Squares (OLS) method executed in Minitab (v. 17) and the genetic programming performed in Eureka Formulize (v. 1.24.0).</p> |
| <p>Keywords:</p> <p>Mathematical Models; Ordinary Least Squares; Genetic Programming.</p> | <p>Theoretical reference: Obtaining a model that mathematically describes the relationship between the independent variable and the response variable is essential to optimizing the process. The model can be described as an approximate representation of the real system or process, while the modeling process is a balance between simplicity and accuracy (X. Chen et al., 2018; Gomes et al., 2019; Sampaio et al., 2022; A. R. S. Silva et al., 2021).</p> |
|  | <p>Method: An Evaluation of the best method for constructing mathematical models was performed using the Adjusted Coefficient of Determination (Radj2) and Akaike's Information Criterion</p> |
| | <p>Results and conclusion: The comparison between the use of the methods showed the superiority of genetic programming over OLS in the construction of mathematical models.</p> <p>Originality/Value: Genetic Programming produces mathematical models that are sometimes differentiated when several replicates are performed, but always with similar explanatory power and with biased characteristic that does not affect in any way the quality of prediction of the dependent variable being studied.</p> |
| | <p>Doi: https://doi.org/10.26668/businessreview/2023.v8i8.3131</p> |

^A PhD in Engineering. Universidade do Estado do Rio de Janeiro (UERJ). Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: nilo.samp@terra.com.br Orcid: <https://orcid.org/0000-0002-6168-785X>

^B Master in Engineering. Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ). Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: jmottareis@gmail.com Orcid: <https://orcid.org/0000-0003-1953-9500>

^C PhD in Engineering. Universidade do Estado do Rio de Janeiro (UERJ). Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: glenio.barros@gmail.com Orcid: <https://orcid.org/0000-0002-6902-599X>

^D PhD in Engineering. Universidade Estadual Paulista (UNESP). Botucatu, São Paulo, Brazil. E-mail: cleginaldo.carvalho@unesp.br Orcid: <https://orcid.org/0000-0001-7364-2096>

^E PhD in Engineering. Universidade de São Paulo (USP). São Paulo, São Paulo, Brazil. E-mail: fmgomes@usp.br Orcid: <https://orcid.org/0000-0001-7694-9835>

^F PhD in Engineering. Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ). Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: luiscesarfmb@gmail.com Orcid: <https://orcid.org/0000-0003-4739-4556>

^G PhD in Engineering. Universidade Estadual Paulista (UNESP). Botucatu, São Paulo, Brazil. E-mail: messias@dequi.eel.usp.br Orcid: <https://orcid.org/0000-0002-8656-0791>

OTIMIZAÇÃO DE RESPOSTAS MÚLTIPLAS: ANÁLISE COMPARATIVA ENTRE MODELOS OBTIDOS PELO MÍNIMO MÉTODO ORDINÁRIO E PROGRAMAÇÃO GENÉTICA

RESUMO

Finalidade: Este trabalho tem como objetivo analisar e comparar o desempenho entre o método dos mínimos quadrados ordinários (OLS) executado no Minitab (v. 17) e a programação genética realizada no Eureka Formulize (v. 1.24.0).

Referência teórica: Obter um modelo que descreva matematicamente a relação entre a variável independente e a variável de resposta é essencial para otimizar o processo. O modelo pode ser descrito como uma representação aproximada do sistema ou processo real, enquanto o processo de modelagem é um equilíbrio entre simplicidade e precisão (X. Chen et al., 2018; Gomes et al., 2019; Sampaio et al., 2022; A. R. S. Silva et al., 2021).

Método: A Avaliação do melhor método para construir modelos matemáticos foi realizada utilizando-se o Coeficiente Ajustado de Determinação (Radj²) e o Critério de Informação de Akaike.

Resultados e conclusão: A comparação entre o uso dos métodos mostrou a superioridade da programação genética sobre o OLS na construção de modelos matemáticos.

Originalidade/Valor: A Programação Genética produz modelos matemáticos que às vezes são diferenciados quando vários replicados são realizados, mas sempre com poder explicativo semelhante e com características tendenciosas que não afetam de forma alguma a qualidade de predição da variável dependente sendo estudada.

Palavras-chave: Modelos Matemáticos, Quadrados Mínimos Ordinários, Programação Genética.

OPTIMIZACIÓN DE RESPUESTA MÚLTIPLE: ANÁLISIS COMPARATIVO ENTRE MODELOS OBTENIDOS POR EL MÉTODO DE MÍNIMOS ORDINARIOS Y PROGRAMACIÓN GENÉTICA

RESUMEN

Objetivo: Este trabajo tiene como objetivo analizar y comparar el desempeño entre el método de mínimos cuadrados ordinarios (MCO) ejecutado en Minitab (v. 17) y la programación genética realizada en Eureka Formulize (v. 1.24.0).

Referencia teórica: La obtención de un modelo que describa matemáticamente la relación entre la variable independiente y la variable respuesta es esencial para optimizar el proceso. El modelo puede describirse como una representación aproximada del sistema o proceso real, mientras que el proceso de modelado es un equilibrio entre simplicidad y precisión (X. Chen et al., 2018; Gomes et al., 2019; Sampaio et al., 2022; A. R. S. Silva et al., 2021).

Método: Se realizó una evaluación del mejor método para la construcción de modelos matemáticos utilizando el coeficiente de determinación ajustado (Radj²) y el criterio de información de Akaike.

Resultados y conclusión: La comparación entre el uso de los métodos mostró la superioridad de la programación genética sobre OLS en la construcción de modelos matemáticos.

Originalidad/Valor: La Programación Genética produce modelos matemáticos que a veces se diferencian cuando se realizan varias réplicas, pero siempre con similar poder explicativo y con características sesgadas que no afectan en modo alguno la calidad de predicción de la variable dependiente que se estudia.

Palabras clave: Modelos Matemáticos, Mínimos Cuadrados Ordinarios, Programación Genética.

INTRODUCTION

Obtaining a model that mathematically describes the relationship between the independent variable and the response variable is essential to optimizing the process. The model can be described as an approximate representation of the real system or process, while the modeling process is a balance between simplicity and accuracy (X. Chen et al., 2018; Gomes et al., 2019; Sampaio et al., 2022; A. R. S. Silva et al., 2021). Empirical models are built based on statistical analysis of experimental observations using regression techniques. Compared to the phenomenological model, the empirical model has the disadvantage of not extrapolating the

data, making the model valid only within the experimental data collection process used to obtain the model (Chau, 2017; V. C. P. Chen et al., 2006; Zhang et al., 2020).

Determining a process improvement is typically complex due to variations in customer demand and technological advances. Generally, several responses must be considered in order to achieve an overall process improvement (Salido et al., 2016). It is important to note that an optimization process does not necessarily imply the determination of optimal operating conditions since it is practically impossible to establish the optimal point due to the large number of variables that impact a process. Instead, what can be determined are improvement conditions from the selection of maximum points within a predetermined search space (Gomes et al., 2019; Ivanov et al., 2016).

Genetic Programming (GP) is part of a broader research field called evolutionary computing, which involves the development of global search and optimization algorithms based on the theory of biological evolution (Garg & Lam, 2015; Katebi et al., 2017; Wan et al., 2018). Ordinary Least Squares (OLS) is a mathematical optimization method that aims to find the best fit for a data set by trying to minimize the sum of squared differences between the estimated and observed data values (Beena & Kumaran, 2010; Cribari-Neto & Lima, 2014; Lakshmi et al., 2021).

The difficulty encountered when working with phenomenological models is to determine all the physical and chemical properties that may affect the process in some way. Many studies applying Design of Experiments (DOE) to optimize processes, especially processes involving multiple responses, have neglected the quality of the model obtained (Ch'ng et al., 2005; Derringer & Suich, 1980; Khuri & Conlon, 1981; Pang et al., 2020; Pinto & Pereira, 2021). This one examines the use of genetic programming to obtain models with higher predictability than those obtained by Ordinary Least Squares (OLS), and the use of models with multiple responses in the optimization process.

LITERATURE REVIEW

In this section, the scientific literature is reviewed to introduce the theoretical foundations of the GP and OLS topics. Papers from journals with high impact factor were prioritized.

Genetic Programming

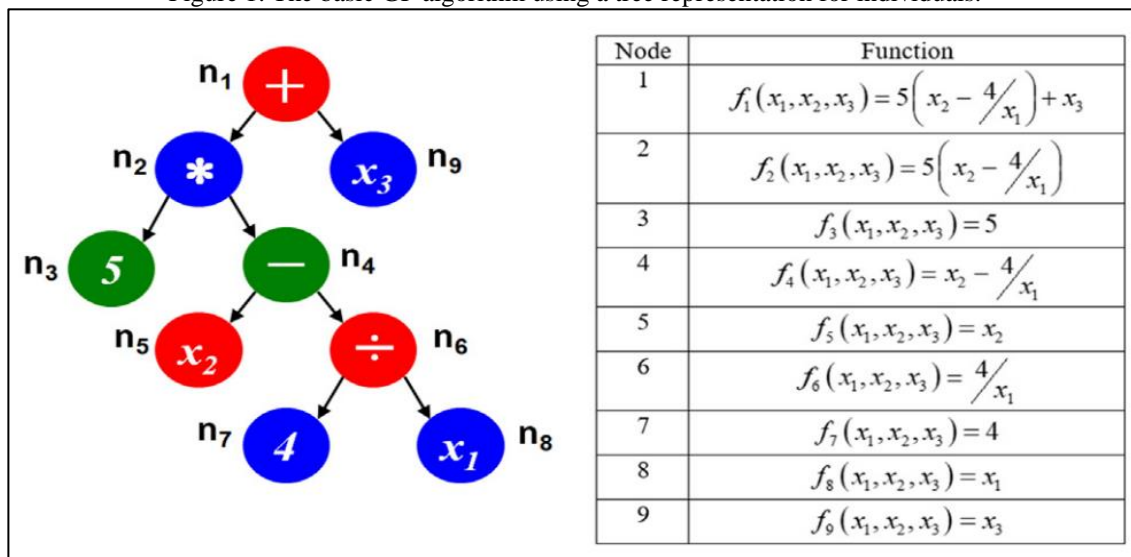
In GP, each chromosome (individual in the population) represents a possible solution to a problem, and is composed of a string of genes. The initial population is taken at random to

serve as a starting point for the algorithm. A fitness function is defined to check the fitness of the chromosome to the environment. Based on the fitness value, chromosomes are selected and crossover and mutation operations are performed on them to produce offspring for the new population. The fitness function evaluates the quality of each offspring. The process is repeated until enough offspring are created (Kalra & Singh, 2015; Wu & Yang, 2013).

GP operates by searching for syntactic expressions that have evolved based on candidate solutions in order to find the expression that best describes the relationship between a set of independent variables and dependent variables Unlike ordinary optimization methods, in conventional optimization methods, potential solutions are represented by numbers (usually vectors of real numbers), while symbolic optimization algorithms represent potential solutions in a structured order of various symbols. One of the most popular methods of structure representation is the binary tree (Wu & Yang, 2013).

A population member in GP is a hierarchically structured tree consisting of functions and terminals. The functions and terminals are selected from a set of functions and a set of terminals, as shown in Figure 1. The set of F operators may contain the basic arithmetic operations. However, it can also include other mathematical methods (Amir Haeri et al., 2017).

Figure 1. The basic GP algorithm using a tree representation for individuals.



Source: Gomes et al. (2019).

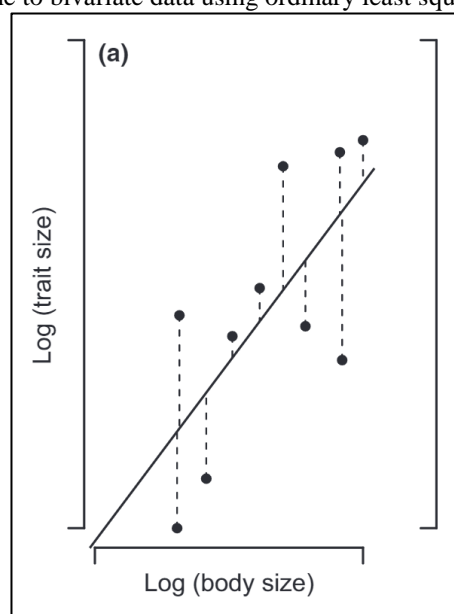
The process of using GP to derive a mathematical model is to generate a series of initial equations that describe the relationship between input variables and output variables. The expression of these equations usually uses the tree form, in which the mathematical parameters and functions that make up the model are expressed in the tree leaves and the answers in the

root(Poli et al., 2008). Compared to other techniques for modeling nonlinear problems, the major advantage of GP is that GP can create models with relatively low errors and does not require the behavior of the dependent and independent variables in the prior process. GP has many applications in modeling problems involving nonlinear equations, emphasizing its greatest use in time series prediction (Wu & Yang, 2013).

Ordinary Least Squares

OLS regression is used to fit a straight line through the data, but has a reputation for underestimating the slope when there is measurement error. Ordinary least squares regression is suitable for bivariate data lines, so for all data points, the vertical (square) distance from the data point to the line is minimized (Figure 2) (Gomes et al., 2019; Kilmer & Rodríguez, 2017; Qasim et al., 2020).

Figure 2. Fitting a line to bivariate data using ordinary least squares (OLS) regression.



Source: Adapted: Kilmer and Rodríguez (2017).

The slope of this line is described by the equation $b_{OLS} = \text{cov}(x,y)/\text{var}(x)$. Therefore, if polycystic ovary changes or the variance of the x-axis variable changes, the OLS slope will change. Since OLS regression uses vertical residuals to fit a line, the values on the horizontal axis are assumed to be perfectly measured, and any deviation from the data points on the regression line is attributed to the variables plotted on the vertical axis (Kilmer & Rodríguez, 2017; Liu & Piantadosi, 2017; Salmeron et al., 2017). However, the classical OLS estimation

used for linear regression is no longer applicable due to the lack of sufficient degrees of freedom (Wang & Leng, 2016).

EVALUATION MODEL CRITERIA

When fitting regression models to experimental data using models with few fitting parameters, they may not be satisfactory in describing the behavior of the data, their estimates are far from the experimentally obtained values, which results in a lack of model fit to the experimental data. Very complex models tend to distribute experimental errors in their parameters, resulting in parameters with low or no statistical significance, and, although they tend to describe the experimental points reliably, they may have poor ability to represent the behavior of the process that generated the data due to the occurrence of random noise in the response (Pitt & Myung, 2002). Therefore, the best model will be the one that aligns a good fit to the experimental data for a good ability to represent the process behavior. To obtain a balanced model, one must use information-theoretic criteria proposed for mathematical model selection, such as the adjusted R^2 and Akaike criterion information.

Adjusted R^2

When comparing models with different numbers of parameters, it is appropriate to use the coefficient of determination adjusted to the number of parameters of each model, so that they are compared under equal conditions and Eq. (Draper & Smith, 1998).

$$Ra^2 = \frac{(n-1).R^2-p}{n-p-1} \quad (1)$$

p = Number of parameters of the regression model
n = Number of observations
 R^2 = Coefficient of Determination

The adjusted R^2 compares the explanatory power of regression models that contain different numbers of predictors. The adjusted R^2 is a modified version of R^2 that has been adjusted for the number of predictors in the model. The adjusted R^2 increases only if a new term added improves the model more than would be expected by chance. It decreases when a predictor improves the model less than expected by chance (Draper & Smith, 1998).

Akaike Information Criterion

Akaike's Information Criterion (AIC) is an asymptotically unbiased opinion estimator of the Kullback-Leibler (K-L) divergence. The K-L divergence can be interpreted as a "divergence" between complete reality and a model. Thus, the best model loses the least information relative to other models in the set (Burnham & Anderson, 2004; Pinto & Pereira, 2021). The (AIC) is obtained from the solution of Eq. (2).

$$AIK = -2 \ln(L) + 2K \quad (2)$$

Where:

L is the maximum point of the Log-Likelihood function, and K is the number of parameters estimated by the model plus one. The first term represents a measure of lack of fit, while the second term represents a measure of model complexity. In the case of ordinary least squares estimation, the (AIC) can be obtained from Eq. (3).

$$AIC = N \cdot \ln \frac{SSN}{N} + 2K \quad (3)$$

Where:

N is the number of points used to obtain the model (sample size). When more parameters are added to a model, the first term becomes smaller while the second term becomes larger. When a statistical model is used to represent a given process, the representation will never be exact, that is, the model will never be perfect and some information will certainly be lost. The AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model and the lower the AIC score.

When N is small compared to K for the largest model in the candidate set (as a rule, $N / K < 40$), it is recommended to use the Akaike information criterion corrected for small samples (AICC) (Burnham & Anderson, 2004). The AICC is given by Eq. (4).

$$AIC + \frac{2K \cdot (K+1)}{N-K-1} \quad (4)$$

The use of the AICC instead of the AIC is preferred, as it is more accurate for small samples and gives very similar results for large samples (Alrubaie et al., 2007). Determining the differences of the AICC (Δ_i) allows for quick comparison and ranking of candidate models. For the i-th model, Δ_i is given by Eq. (5).

$$\Delta_i = AICci - \min AICc \quad (5)$$

Where:

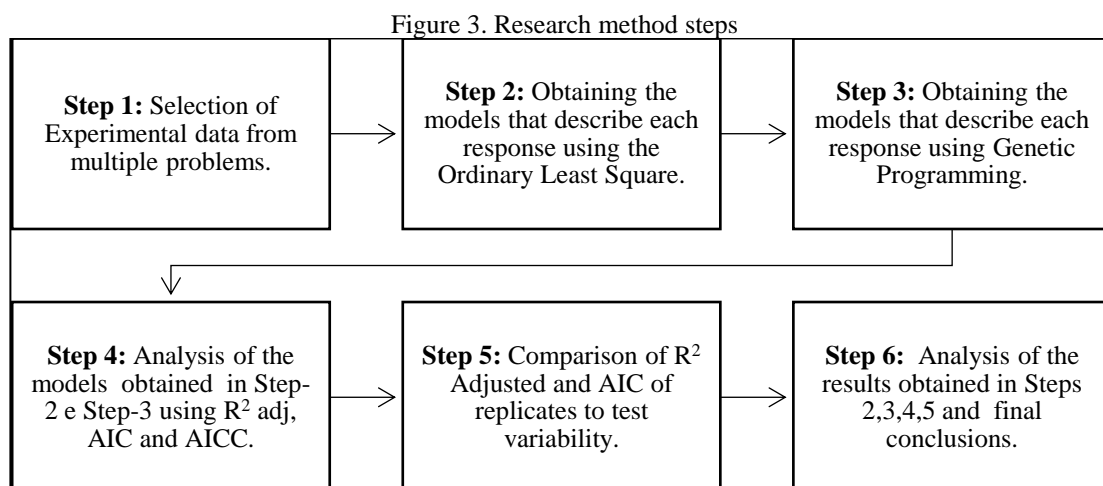
min AICC is the lowest AICC value among all the models evaluated. The Δ_i of the best generated model is equal to zero, while the rest of the models have positive values, and the higher the value of Δ_i for the model, the worse the quality of its fit will be. As a rule, models with $\Delta_i \leq 2$ have substantial predictability support, those with $2 < \Delta_i \leq 7$ have considerably less support, and models with $\Delta_i > 7$ have no support (Burnham & Anderson, 2004).

Homoscedasticity

In Statistics, a sequence of random variables is homoscedastic if all its random variables have the same finite variance. This is also known as homogeneity of variance. Assuming that a variable is homoscedastic when in fact it is heteroscedastic results in unbiased but inefficient point estimates and biased standard error estimates, and can result in overestimation of the quality of fit as measured by Pearson's coefficient (Diniz et al., 2012; Gomes et al., 2019; Gonçalves & Ghosh, 2022; Wilcox, 2007).

RESEARCH METHOD

This work can be classified as an applied research because it aims at providing improvements in the current literature, with normative empirical objectives, aiming at developing policies and strategies that improve the current situation (Araujo et al., 2021; H. de O. G. da Silva et al., 2021; Will M. Bertrand & Fransoo, 2002). The problem approach is quantitative, as is the modeling and simulation research method. The research steps were performed following the sequence shown in Figure 3.



Source: Authors (2023).

Step 1: The Experimental data were selected from the work of Naderi and Khomehchi (2017); Sathiya et al, (2011). This choice was based on the fact that this work is the

most widely used data source for comparing optimization methods with multiple responses.

Step 2: Models were generated from the experimental data, describing the previously selected responses, using the Ordinary Least Squares technique with the help of Minitab v.19; the models were refined with R^2_{adj} (Adjusted Coefficient of Determination).

Step 3: Models were generated from the experimental data, describing the previously selected responses, using the Symbolic Regression Technique via Genetic Programming with the help of Eureqa v 1.24.0; a more detailed review of this software can be found at (Dubčáková, 2011).

Step 4: An analysis of the models obtained in steps 2 and 3 was performed, comparing values of adjusted R^2 , AIC and AICC. With this comparison, the models can be classified according to their higher predictability.

Step 5: During Step 3, three replicates of each model were run, and the comparison of the adjusted R^2 and AIC of the replicates were compared with each other to show that the variability is very small and the models are reliable.

Step 6: The conclusions presented at the end of this paper were drawn from the results obtained in the previous steps.

Case 1

The problem described by Naderi and Khamehchi (2017) is the optimization of Recovery Factor (RF) and Cumulative Water Production (LnWp). The present study has two objectives. The first objective is to find the optimal number of water wells and their corresponding locations. The second objective is to find the optimal production rate and its corresponding locations, the drilling thickness, and the lower bound of the pipe head pressure using a new stick-inspired metaheuristic algorithm. In this regard, DOE via Response Surface and Symbolic Regression (RSM) via Genetic Programming were used to develop equations to model RF and Wp, and compare the fit of both methods.

DOE is a systematic method for obtaining the most information by running a minimum set of experiments in order to determine the input-output relationship. Table 1 shows the seven independent variables and the two dependent variables. RSM is a collection of mathematical and statistical techniques for obtaining the relationships between independent and dependent variables (response) (Box & Wilson, 1951).

The Table shows the Dependent Variables LnWp and RF, and the seven independent variables (X1, X2, X3, X4, X5, X6 and X7).

Table 1. Dependent and Independent Variables

| | |
|------|--|
| X1 | Coded variable for well location I. |
| X2 | Variable coded for the location of well J. |
| X3 | Variable coded for average reservoir permeability. |
| X4 | Variable coded for permeability anisotropy. |
| X5 | Variable coded for gas production rate. |
| X6 | Variable coded for borehole thickness |
| X7 | Variable coded for pipe head pressure |
| Wp | Cumulative Water Production. |
| LnWp | Cumulative Water Production. |
| RF | Gas Recovery Factor (%) |

Source: Authors (2023).

Table 2 shows the results of all 57 different BBD-based flow simulations for seven factors with three levels that were performed using the Eclipse 100 black oil simulator.

Table 2. Result of 57 flow simulations based on BBD

| # | X ₁ | X ₂ | X ₃ | X ₄ | X ₅ | X ₆ | X ₇ | RF% | Ln Wp |
|----|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-------|-------|
| 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 51.03 | 10.73 |
| 2 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 34.67 | 12.28 |
| 3 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 74.84 | 8.11 |
| 4 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 82.16 | 8.31 |
| 5 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 78.15 | 8.04 |
| 6 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 80.65 | 8.10 |
| 7 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 62.34 | 13.88 |
| 8 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 79.38 | 12.87 |
| 9 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 56.94 | 11.86 |
| 10 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 65.81 | 7.65 |
| 11 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 70.99 | 8.02 |
| 12 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 65.77 | 13.20 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 75.92 | 8.03 |
| 14 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 66.63 | 13.10 |
| 15 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 85.58 | 8.33 |
| 16 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 62.27 | 7.46 |
| 17 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 51.13 | 10.70 |
| 18 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 82.16 | 8.31 |
| 19 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 73.21 | 8.18 |
| 20 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 73.11 | 8.17 |
| 21 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 79.36 | 12.87 |
| 22 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 71.42 | 12.10 |
| 23 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 62.91 | 7.51 |
| 24 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 62.87 | 7.50 |
| 25 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 62.32 | 7.46 |
| 26 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 51.23 | 10.75 |
| 27 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 52.55 | 12.74 |
| 28 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 78.20 | 8.04 |
| 29 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 50.57 | 13.85 |
| 30 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 80.69 | 8.10 |
| 31 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 88.21 | 8.40 |
| 32 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 77.28 | 8.11 |
| 33 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 73.04 | 8.17 |
| 34 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 71.20 | 12.09 |

| | | | | | | | | | |
|----|---|---|---|---|---|---|---|-------|-------|
| 35 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 59.87 | 7.68 |
| 36 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 51.11 | 10.70 |
| 37 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 63.85 | 7.69 |
| 38 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 77.29 | 8.11 |
| 39 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 87.60 | 8.67 |
| 40 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 52.34 | 12.73 |
| 41 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 56.83 | 11.87 |
| 42 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 76.33 | 8.83 |
| 43 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 76.39 | 8.26 |
| 44 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 66.70 | 7.87 |
| 45 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 63.77 | 7.69 |
| 46 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 76.75 | 7.97 |
| 47 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 49.85 | 13.81 |
| 48 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 76.74 | 7.97 |
| 49 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 76.46 | 8.90 |
| 50 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 47.26 | 13.69 |
| 51 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 67.13 | 7.16 |
| 52 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 49.61 | 13.81 |
| 53 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 76.00 | 8.40 |
| 54 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 72.57 | 9.45 |
| 55 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 85.61 | 8.33 |
| 56 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 73.06 | 8.17 |
| 57 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 65.80 | 7.65 |

Source: Authors (2023).

Table 4 shows the coefficient of determination (R^2) and adjusted coefficient of determination (R^{2adj}) for the proxy equation of the recovery factor and cumulative water yield. These statistics are used to show the quality of fit for regressions. R^2 is a number between zero and one that shows how well the well data fits a statistical model. In other words, it measures the percentage of the variability in the process explained by the fitted model (Naderi & Khamehchi, 2017).

Case 2

The problem described by Sathiya et al. (2011) is the use of Taguchi's method that determines the optimal results of finite analytical data and the dominant factors involved in the optimization of laser welding from finite analytical data. In this study, three level process parameters, i.e. beam power (BP), travel speed (TS) and focal position (FP) are considered. In this study, an orthogonal L27 matrix with 26 degrees of freedom was used. Twenty-seven experiments on the three shielding gases are required to study the entire welding parameter space when the orthogonal L27 arrangement is used. The experimental results of the laser weld, i.e. tensile strength, bead width, and depth of penetrations, are shown in Table 3. The weld profiles were obtained by sectioning and polishing with suitable abrasive and diamond paste. The weld samples were etched with 10% oxalic acid, an electrolyte, to indicate and increase the contrast of the fusion zone with the base metal.

The joint quality is evaluated by studying X1= BW (Weld Bead Geometry) X2= TST (Tensile Strength), and X3= (PDO) Penetration Pitting using Argon, Nitrogen and Helium gases. The goal is to model $R1=TST(BP,TS,FP)Arg$, $R2=BW(BP,TS,FP)Arg$, $R3=DOP(BP,TS,FP)Arg$, $R4=TST(BP,TS,FP)Nit$, $R5=BW(BP,TS,FP)Nit$, $R6=DOP(BP,TS,FP)Nit$, $R7=TST(BP,TS,FP)Hel$, $R8=BW(BP,TS,FP)Hel$, $R9=DOP(BP,TS,FP)Hel$.

Table 3. Experimental results

| Beam power | | Focal position | Argon | | | Nitrogen | | | Helium | | |
|--------------|----------|----------------|-----------|---------|----------|-----------|---------|----------|-----------|---------|----------|
| Travel speed | in m/min | | TST (Mpa) | BW (mm) | DOP (mm) | TST (Mpa) | BW (mm) | DOP (mm) | TST (Mpa) | BW (mm) | DOP (mm) |
| 1 | 1 | 1 | 615 | 2.032 | 2.891 | 589 | 1.632 | 2.699 | 620 | 1.251 | 2.854 |
| 1 | 1 | 2 | 622 | 2.041 | 2.902 | 590 | 1.629 | 2.696 | 618 | 1.249 | 2.859 |
| 1 | 1 | 3 | 620 | 2.043 | 2.923 | 585 | 1.634 | 2.697 | 615 | 1.250 | 2.858 |
| 1 | 2 | 1 | 589 | 2.109 | 2.722 | 612 | 1.579 | 2.719 | 627 | 1.152 | 2.946 |
| 1 | 2 | 2 | 585 | 2.112 | 2.693 | 605 | 1.582 | 2.712 | 622 | 1.149 | 2.942 |
| 1 | 2 | 3 | 579 | 2.113 | 2.732 | 609 | 1.580 | 2.716 | 624 | 1.148 | 2.947 |
| 1 | 3 | 1 | 568 | 1.729 | 2.895 | 601 | 1.299 | 2.770 | 630 | 1.227 | 2.660 |
| 1 | 3 | 2 | 565 | 1.732 | 2.932 | 598 | 1.301 | 2.772 | 629 | 1.230 | 2.664 |
| 1 | 3 | 3 | 570 | 1.736 | 2.879 | 603 | 1.303 | 2.769 | 634 | 1.229 | 2.662 |
| 2 | 1 | 1 | 610 | 1.570 | 2.893 | 570 | 1.451 | 2.810 | 612 | 1.232 | 2.780 |
| 2 | 1 | 2 | 612 | 1.569 | 2.821 | 569 | 1.449 | 2.812 | 609 | 1.230 | 2.782 |
| 2 | 1 | 3 | 605 | 1.563 | 2.851 | 572 | 1.450 | 2.811 | 605 | 1.231 | 2.781 |
| 2 | 2 | 1 | 623 | 1.959 | 2.693 | 565 | 1.629 | 2.790 | 620 | 1.331 | 2.795 |
| 2 | 2 | 2 | 625 | 1.961 | 2.714 | 570 | 1.632 | 2.786 | 625 | 1.330 | 2.791 |
| 2 | 2 | 3 | 619 | 1.960 | 2.737 | 557 | 1.630 | 2.789 | 619 | 1.332 | 2.796 |
| 2 | 3 | 1 | 621 | 2.119 | 2.735 | 618 | 1.920 | 2.659 | 629 | 1.273 | 2.785 |
| 2 | 3 | 2 | 619 | 2.120 | 2.694 | 620 | 1.917 | 2.664 | 635 | 1.275 | 2.781 |
| 2 | 3 | 3 | 616 | 2.123 | 2.727 | 624 | 1.922 | 2.663 | 633 | 1.272 | 2.780 |
| 3 | 1 | 1 | 620 | 1.646 | 2.723 | 629 | 1.240 | 2.453 | 594 | 1.439 | 2.887 |
| 3 | 1 | 2 | 625 | 1.650 | 2.745 | 625 | 1.242 | 2.451 | 591 | 1.438 | 2.883 |
| 3 | 1 | 3 | 619 | 1.642 | 2.731 | 622 | 1.239 | 2.452 | 590 | 1.434 | 2.885 |
| 3 | 2 | 1 | 629 | 1.610 | 2.787 | 635 | 1.479 | 2.859 | 621 | 1.362 | 2.631 |
| 3 | 2 | 2 | 633 | 1.609 | 2.725 | 639 | 1.470 | 2.855 | 619 | 1.360 | 2.632 |
| 3 | 2 | 3 | 631 | 1.606 | 2.753 | 641 | 1.473 | 2.857 | 620 | 1.361 | 2.633 |
| 3 | 3 | 1 | 632 | 1.828 | 2.732 | 599 | 1.619 | 2.830 | 573 | 1.329 | 2.870 |
| 3 | 3 | 2 | 635 | 1.831 | 2.794 | 602 | 1.612 | 2.831 | 579 | 1.330 | 2.873 |
| 3 | 3 | 3 | 630 | 1.829 | 2.721 | 605 | 1.617 | 2.829 | 575 | 1.331 | 2.875 |

Source: Authors (2023).

RESULTS AND DISCUSSION

Case 1

The models obtained using Ordinary Least Squares, using Minitab v.17, are shown in Eqs. (6)-(7).

$$\text{Lnwp} = 7.266 - 0.010 X_1 + 0.008 X_2 - 0.868 X_3 + 1.750 X_4 + 0.048 X_5 + 1.963 X_6 - 0.418 X_7 + 1.536 X_1 * X_1 + 0.824 X_2 * X_2 + 0.672 X_3 * X_3 + 1.307 X_4 * X_4 + 1.254 X_6 * X_6 + 0.003 X_1 * X_5 - 0.003 X_1 * X_5 + 0.003 X_2 * X_6 - 1.400 X_3 * X_4 + 0.856 X_4 * X_6 \quad (6)$$

$$\text{RF} = 68.13 + 2.27 X_1 + 0.49 X_2 - 1.39 X_3 + 1.19 X_4 + 1.38 X_5 + 6.00 X_6 + 2.24 X_7 - 4.53 X_1 * X_7 - 4.86 X_2 * X_7 - 6.95 X_3 * X_4 \quad (7)$$

The models proposed by the Genetic Programming method with the three replicates using the Eureqa Formulize software are found in Eqs. (8) - (13).

$$\text{Ln}(Wp)1 = X_4 + X_4.X_6 + X_3.X_2^2 - \frac{49.8}{X_6 + X_4.X_2^2 - 4.82} - X_3 - X_3.X_4 - X_5^2 - X_7^2 - X_6.X_5^2 - X_2^2.X_3^2 - 0.424.X_7 \quad (8)$$

$$\text{RF1} = 68.1 + 6.X_2 + 6.X_6 + 6.X_4.X_2^2 - 6.X_3.X_4 - 11X_5.X_6 - 9.71.X_1.X_2.X_4 - 13.8.X_2.X_3^2 \quad (9)$$

$$\text{Ln}(Wp)2 = 9.56 + X_4 + 2.3.X_5 + X_4.X_5 + X_2^2 + 1.82.X_4.X_2^2 - X_3 - X_3.X_4 - X_5.X_4^2 - 2.56.X_2^2.X_7^2 \quad (10)$$

$$\text{RF2} = 68.1 + 5.47.X_2 + 5.47.X_5 + 2.28X_1 - 7.X_3.X_4 - 10.4.X_5.X_6 - 10.4.X_1.X_2.X_4 - 13.3.X_2.X_3^2 \quad (11)$$

$$\text{Ln}(Wp)3 = 9.75 + X_4 + 2.31.X_5 + X_4.X_5 + 1.78.X_4.X_2^2 + X_2^2.X_4^2 - X_3 - X_5.X_4^2 - 1.42.X_3.X_4 - 1.78.X_2.X_7^2 \quad (12)$$

$$\text{RF3} = 68.1 + 5.29.X_5 + 5.29.X_2.X_6 - 7.33.X_3.X_4 - 11.4.X_5.X_6 - 9.55.X_1.X_2.X_4 - 13.1.X_2.X_3^2 \quad (13)$$

The values of R^2 , R^2_{adj} , and $R^2 - R^2_{adj}$ of the models presented in equations (6) to (13) are shown in Table 4, and it is possible to see that the R^2_{adj} is always higher in the model obtained by Genetic Programming than in the one obtained by the Ordinary Least Squares Method, showing that the Genetic Programming optimization is the best method for building mathematical models.

Table 4. Comparison of the models based on the results of R^2_{adj}

| | R^2 | R^2_{adj} | $R^2 - R^2_{adj}$ |
|------------------|--------|-------------|-------------------|
| Ln (Wp)MT | 89.22% | 82.75% | 6.47% |
| RF1MT | 24.37% | 7.93% | 16.44% |
| Ln (Wp)1E | 94.69% | 94.59% | 0,1% |
| Ln (Wp)2E | 90.60% | 90.07% | 0,53% |
| Ln (Wp)3E | 91.54% | 90.89% | 0,65% |
| RF1E | 51.45% | 44.52% | 6,93% |
| RF2E | 48.68% | 40.99% | 7.69% |
| RF3E | 48.36% | 42.17% | 6.19% |

Source: Authors (2023).

The AIC and AICC values for the models presented in equations (6) to (13) are shown in Table 5, and it can be seen that the AICC (since the rule, $N/K < 40$) for all models is always lower in the model obtained by Genetic Programming than that obtained by the Ordinary Least Squares Method, showing that Genetic Programming optimization is the best method for building mathematical models.

Table 5. Comparison of the models based on the results of the AIC

| Modelo | AIC | AICC |
|-----------------------------|------------|-------------|
| Ln (W_p)MT | -1.6678 | 14.02451 |
| RF1MT | 286.7189 | 291.5015 |
| Ln (W_p)1E | -74.67 | -74.5973 |
| Ln (W_p)2E | 8.043 | 8.49583 |
| Ln (W_p)3E | -35.8763 | -35.1071 |
| RF1E | 256.198 | 258.4837 |
| RF2E | 259.0617 | 261.3474 |
| RF3E | 265.629 | 267.9147 |

Source: Authors (2023).

The AICC and Δ values of the models presented in equations (6) to (13) appear in Table 6, for the replications made by Genetic Programming the average AICC value of the three values was calculated. The Δ is calculated by decreasing the smallest value of it by itself and the smallest value, by the largest value, a value greater than and equal to 7 was found, which means that the Ordinary Least Squares Method is the worst method for building mathematical models.

Table 6. Comparison of the models based on the results of the Δ

| Modelo | AICC | Δ |
|----------------------------------|-------------|----------------------------|
| Ln (W_p)MT | 14.02451 | 47.7607 |
| Média (Ln(W_p)) | -33.7362 | 0 |
| RF1MT | 291.5015 | 28.9196 |
| Média (RF) | 262.5819 | 0 |

Source: Authors (2023).

Since the worst AICC value of the models obtained by Genetic Programming is still better (a smaller value) than the value obtained by the Ordinary Least Squares Method, the best value obtained by Genetic Programming was chosen to perform the Normality Test to verify Homoscedasticity. The values presented in Table 7 show that model 2 is Homoscedastic because the p-value is greater than 0.05 (the distribution of the residuals is normal when p-value

is greater than 0.05). While model 1 is Heteroscedastic because its p-value is less than 0.05, so only model 2 is unbiased.

Table 7. Results of Homoscedasticity test

| Modelo | p-Value | p-Value < 0,05 (Homoscedastic) |
|------------------------------|---------|--------------------------------|
| Ln (W _p) – Rep-1 | 0,006 | Heteroscedastic |
| RF-Rep-1 | 0,874 | Homoscedastic |

Source: Authors (2023).

Case 2

The models obtained using Ordinary Least Squares, using Minitab v.17, are shown in Eqs. (14)-(22).

$$R1 = 611.74 + 18.94X1 - 0.33X2.X3 \quad (14)$$

$$R2 = 1.8460 - 0.1331.X1 + 0.0717X2 + 0.0007X3 + 0.1224.X1.X2 \quad (15)$$

$$R3 = 2.7831 - 0.0477.X1 \quad (16)$$

$$R4 = 601.63 + 10.83.X1 + 6.61.X2 - 9.X1.X2 \quad (17)$$

$$R5 = 1.5381 + 0.1766.X1.X2 \quad (18)$$

$$R6 = 2.7315 - 0.0074.X1 + 0.0503.X2 + 0.0763.X1.X2 \quad (19)$$

$$R7 = 613.63 - 14.28.X1 + 3.5.X2 - 7.33.X1.X2 \quad (20)$$

$$R8 = 1.28796 + 0.0833.X1 - 0.0214.X1.X2 \quad (21)$$

$$R9 = 2.8012 - 0.0124.X1 - 0.0344.X2 + 0.0005.X3 + 0.0457.X1.X2 + 0.0014.X1.X2.X3 \quad (22)$$

The models proposed by the Genetic Programming method with the three replicates using the Eureka Formulize software are found in Eqs. (23) - (49).

$$(R1)1 = 618 + 23.7X1 + 4.83.X2 + 15.6.X1.X2 - X3 - 2.56X3^2 - 7.07.X1.X2^2 - 14.9X2.X1^2 \quad (23)$$

$$(R2)1 = 1.9 + 0.276.X2 + 0.124.X1.X2 + 0.177.X1.X2^2 - 0.252.X1 - 0.0369.X1^2 - 0.0519.X2^2 - 0.307.X2.X1^2 \quad (24)$$

$$(R3)1 = 2.71 + 0.0197.X1 + 0.00483.X1.X2 + 0,0821X2^2 + 0.0306.X1^2 + 0.0713.X2.X1^2 - 0.0682.X2 - 0.101.X1.X2^2 \quad (25)$$

$$(R4)1 = 564 + 25.2.X2 + 10.8.X1 + 57.8X1^2 + 31.5X2^2 - 9.X1.X2 - 27.8.X2.X1^2 - 49.3X1^2.X2^2 \quad (26)$$

$$(R5)1 = 1.63 + 0.235.X2 + 0.177.X1.X2 + 0.0547.X2^2 - 0.0304.X1 - 0.103.X1^2 - 0.224.X2.X1^2 - 0.135.X1^2.X2^2 \quad (27)$$

$$(R6)1 = 2.8 + 0.0727.X1 + 0.0727.X1.X2 + 0.185.X2.X1^2 - 0.0727.X2 - 0.119.X1.X2^2 - 0.0263.X1^2.X2^2 \quad (28)$$

$$(R7)1 = 621 + 11.8X2 + 2.17.X2.X3 - X1 - 7.33.X1.X2 - 12.5.X2.X1^2 - 19.3.X1.X2^2 - 17.3.X1^2.X2^2 \quad (29)$$

$$(R8)1 = 1.33 + 0.0735.X1 + 0.021.X2 + 0.121.X1^2.X2^2 - 0.0215.X1.X2 - 0.0616.X1^2 - 0.079X2^2 - 0,0535.X2.X1^2 \quad (30)$$

$$(R9)1 = 2.79 + 0.046.X1.X2 + 0.216.X1.X2^2 + 0,0385.X1^2.X2^2 - 0.157.X1 - 0.00899.X2^2 - 0.052.X2.X1^2 \quad (31)$$

$$(R1)2 = 618 + 23.3.X1 + 5.17.X2 + 15.4.X1.X2 - X3 - 2.56.X3^2 - 7.39.X1^2 - 6.58.X1.X2^2 - 15.4.X2.X1^2 \quad (32)$$

$$(R2)2 = 1.93 + 0.277.X2 + 0.122.X1.X2 + 0.178.X1.X2^2 - 0.251.X1 - 0,0548.X1^2 - 0.0706.X2^2 - 0.307.X2.X1^2 \quad (33)$$

$$(R3)2 = 2.71 + 0.0197.X1 + 0.00488.X1.X2 + 0.0821.X2^2 + 0.0306.X1^2 + 0.0713.X2.X1^2 - 0.0682.X2 - 0.101.X1.X2^2 \quad (34)$$

$$(R4)2 = 564 + 25.2.X2 + 10.8.X1 + 57.8.X1^2 + 31.5.X2^2 - 9.X1.X2 - 27.8.X2.X1^2 - 49.3.X1^2.X2^2 \quad (35)$$

$$(R5)2 = 1.63 + 0.235.X2 + 0.177.X1.X2 + 0.0545.X2^2 - 0.0304.X1 - 0.103.X1^2 - 0.224.X2.X1^2 - 0.134.X1^2.X2^2 \quad (36)$$

$$(R6)2 = 2.81 + 0.0708.X1 + 0.0763.X1.X2 + 0.187.X2.X1^2 - 0.0746.X2 - 0.0335.X1^2 - 0.0832.X2^2 - 0.117.X1.X2^2 \quad (37)$$

$$(R7)2 = 621 + 11.8.X2 + 2.17.X2.X3 - X1 - 7.33.X1.X2 - 12.5.X2.X1^2 - 19.3.X1.X2^2 - 17.3.X1^2.X2^2 \quad (38)$$

$$(R8)2 = 1.33 + 0.0834.X1 + 0.0213.X2 + 0.133.X1^2.X2^2 - 0.021.X1.X2 - 0.074.X1^2 - 0.0776.X2^2 - 0.0534.X2.X1^2 \quad (39)$$

$$(R9)2 = 2.79 + 0.0457.X1.X2 + 0.216.X1.X2^2 + 0.0377.X1^2.X2^2 - 0.156.X1 - 0.00833X2^2 - 0.0518.X2.X1^2 \quad (40)$$

$$(R1)3 = 618 + 23.7.X1 + 4.83.X2 + 15.6.X1.X2 - X3 - 2.55.X3^2 - 7.07X1^2 - 7.07.X1.X2^2 - 14.9.X2.X1^2 \quad (41)$$

$$(R2)3 = 1.93 + 0.277.X2 + 0.122.X1.X2 + 0.178.X1.X2^2 - 0.251.X1 - 0.0549.X1^2 - 0.0708.X2^2 - 0.307.X2.X1^2 \quad (42)$$

$$(R3)3 = 2.71 + 0.0197.X1 + 0.00483.X1.X2 + 0.0821.X2^2 + 0,0306.X1^2 + 0,0713.X2.X1^2 - 0.0682.X2 - 0.101.X1.X2^2 \quad (42)$$

$$(R4)3 = 564 + 25.2.X2 + 10.8.X1 + 57.8.X1^2 + 31.5.X2^2 - 9.X1.X2 - 27.8.X2.X1^2 - 49.3.X1^2.X2^2 \quad (43)$$

$$(R5)3 = 1.63 + 0.235.X2 + 0.177.X1.X2 + 0.0542.X2^2 - 0.0304.X1 - 0.103.X1^2 - 0.224.X2.X1^2 - 0.134.X1^2.X2^2 \quad (44)$$

$$(R6)3 = 2.79 + 0.0707.X1 + 0.0757.X1.X2 + 0.188.X2.X1^2 - 0.0757.X2 - 0.0505.X2^2 - 0.117.X1.X2^2 - 0.0491.X1^2.X2^2 \quad (45)$$

$$(R7)3 = 621 + 11.8.X2 + 2.17.X2.X3 - X1 - 7.33.X1.X2 - 12.5.X2.X1^2 - 19.3.X1.X2^2 - 17.3.X1^2.X2^2 \quad (46)$$

$$(R8)3 = 1.33 + 0.0833.X1 + 0,0212.X2 + 0.132.X1^2.X2^2 - 0.0214.X1.X2 - 0.073.X1^2 - 0.0766X2^2 - 0.0532.X2.X1^2 \quad (47)$$

$$(R9)3 = 2.79 + 0.0457.X1.X2 + 0.216.X1.X2^2 + 0.0377.X1^2.X2^2 - 0.156.X1 - 0.00882.X2^2 - 0.0518.X2.X1^2 \quad (48)$$

The values of R^2 , R^2_{adj} , and $R^2 - R^2_{adj}$ of the models presented in equations (14) to (49) are shown in Table 8, and again it is always possible to see that the R^2_{adj} is always higher in the model obtained by Genetic Programming than in the one obtained by the Ordinary Least Squares Method, showing that the Genetic Programming optimization is the best method for building mathematical models.

Table 8. Response functions statistics-1

| | R^2 | R^2_{adj} | $R^2 - R^2_{adj}$ |
|---------|--------|-------------|-------------------|
| (R1) MT | 55.74% | 52.06% | 3.68% |
| (R2) MT | 51.42% | 42.06% | 9.36% |
| (R3) MT | 24.21% | 21.18% | 3.03% |
| (R4) MT | 27.10% | 13.85% | 13.25% |
| (R5) MT | 38.08% | 32.93% | 5.15% |
| (R6) MT | 32.08% | 23.22% | 8.86% |
| (R7) MT | 52.59% | 46.40% | 6.19% |
| (R8) MT | 73.99% | 71.82% | 2.17% |
| (R9) MT | 19.30% | 0.09% | 19.21% |
| (R1)1 | 97.41% | 96.63% | 0.78% |
| (R2)1 | 98.98% | 98.61% | 0.37% |
| (R3)1 | 92.35% | 89.53% | 2.82% |
| (R4)1 | 97.73% | 96.89% | 0.84% |
| (R5)1 | 99.52% | 99.34% | 0.18% |
| (R6)1 | 99.69% | 99.59% | 0.1% |
| (R7)1 | 98.92% | 98.60% | 0.32% |
| (R8)1 | 95.76% | 94.20% | 1.56% |
| (R9)1 | 99.94% | 99.92% | 0.02% |
| (R1)2 | 97.41% | 96.46% | 0.95% |
| (R2)2 | 99.45% | 99.25% | 0.2% |

| | | | |
|-------|--------|--------|-------|
| (R3)2 | 92.35% | 89.53% | 2.82% |
| (R4)2 | 97.73% | 96.89% | 0.84% |
| (R5)2 | 99.52% | 99.33% | 0.19% |
| (R6)2 | 99.17% | 98.85% | 0.32% |
| (R7)2 | 98.92% | 98.60% | 0.32% |
| (R8)2 | 97.74% | 96.45% | 1.29% |
| (R9)2 | 99.94% | 99.92% | 0.02% |
| (R1)3 | 97.40% | 96.46% | 0.94% |
| (R2)3 | 99.45% | 99.25% | 0.2% |
| (R3)3 | 92.35% | 89.53% | 2.82% |
| (R4)3 | 97.73% | 96.89% | 0.84% |
| (R5)3 | 99.52% | 99.33% | 0.19% |
| (R6)3 | 99.97% | 99.96% | 0.01% |
| (R7)3 | 98.92% | 98.60% | 0.32% |
| (R8)3 | 97.41% | 96.45% | 0.96% |
| (R9)3 | 99.94% | 99.92% | 0.02% |

Source: Authors (2023).

The AIC and AICC values for the models presented in equations (14) to (49) are shown in Table 9, and it is possible to see that the AICC for all models is always lower in the model obtained by Genetic Programming than in the one obtained by the Ordinary Least Squares Method, showing that the Genetic Programming optimization is the best method for building mathematical models.

Table 9. Response functions statistics-2

| Modelo | AIC | AICC |
|---------|----------|----------|
| (R1) MT | 145.6686 | 146.1686 |
| (R2) MT | -96.7054 | -94.8872 |
| (R3) MT | -142.492 | -142.332 |
| (R4) MT | 166.7845 | 167.828 |
| (R5) MT | -100.408 | -100.248 |
| (R6) MT | -120.853 | -119.81 |
| (R7) MT | 141.5464 | 142.5899 |
| (R8) MT | -168.224 | -167.724 |
| (R9) MT | -121.776 | -118.919 |
| (R1)1 | 114.1467 | 118.3467 |
| (R2)1 | -196.262 | -190.367 |
| (R3)1 | -192.179 | -186.284 |
| (R4)1 | 81.1275 | 87.02224 |
| (R5)1 | -219.308 | -213.413 |
| (R6)1 | -133.384 | -129.184 |
| (R7)1 | 46.141 | 50.341 |
| (R8)1 | -207.732 | -201.837 |
| (R9)1 | -312.175 | -307.975 |
| (R1)2 | 79.3555 | 85.25024 |
| (R2)2 | -211.674 | -205.779 |
| (R3)2 | -192.202 | -186.307 |
| (R4)2 | 81.1275 | 87.02224 |
| (R5)2 | -221.432 | -215.537 |
| (R6)2 | -231.477 | -225.582 |
| (R7)2 | 46.141 | 50.341 |
| (R8)2 | -220.524 | -214.629 |

| | | |
|-------|----------|----------|
| (R9)2 | -312.977 | -308.777 |
| (R1)3 | 79.0965 | 84.99124 |
| (R2)3 | -211.674 | -205.779 |
| (R3)3 | -192.181 | -186.286 |
| (R4)3 | 81.1275 | 87.02224 |
| (R5)3 | -219.314 | -213.419 |
| (R6)3 | -290.625 | -284.73 |
| (R7)3 | 46.141 | 50.341 |
| (R8)3 | -220.352 | -214.457 |
| (R9)3 | -312.977 | -308.777 |

Source: Authors (2023).

The values of the AICC and Δ of the models presented in equations (14) to (49) appear in Table 10; for the replications made by Genetic Programming, the value of the average of the AICC of the three values was calculated. The Δ is calculated by decreasing the smallest value of it itself and the smallest value of the largest value, a value greater than and equal to 7 was found which means that the Ordinary Least Squares Method is the worst method for building mathematical models.

Table 10. Comparison of the models based on the results of the Δ

| Modelo | AICC | Δ |
|-------------------|-------------|----------------------------|
| (R1) MT | 146.1686 | 49.973 |
| Média das Rep(R1) | 96.196 | 0 |
| (R2) MT | -94.8872 | 105.7528 |
| Média das Rep(R2) | -200.64 | 0 |
| (R3) MT | -142.332 | 43.958 |
| Média das Rep(R3) | -186.29 | 0 |
| (R4) MT | 167.828 | 80.8058 |
| Média das Rep(R4) | 87.0222 | 0 |
| (R5) MT | -100.248 | 113.875 |
| Média das Rep(R5) | -214.123 | 0 |
| (R6) MT | -119.81 | 93.355 |
| Média das Rep(R6) | -213.165 | 0 |
| (R7) MT | 142.5899 | 92.2489 |
| Média das Rep(R7) | 50.341 | 0 |
| (R8) MT | -167.724 | 42.584 |
| Média das Rep(R8) | -210.308 | 0 |
| (R9) MT | -118.919 | 189.591 |
| Média das Rep(R9) | -308.51 | 0 |

Source: Authors (2023).

The values presented in Table 11 show that only model 2 is Heteroscedastic (p-value < 0.05); therefore, only this model is biased. It is interesting to note that verifying Homoscedasticity does not mean that the quality of the model is low or high, but rather that there is or is not a tendency to overestimate or underestimate the predicted values.

Table 11. Results of Homoscedasticity test

| Modelo | p-Value | p-Value < 0,05 (Homoscedastic) |
|----------|---------|-----------------------------------|
| R1-Rep-3 | 0,527 | Homoscedastic |
| R2-Rep-2 | < 0,005 | Heteroscedastic |
| R3-Rep-2 | 0,861 | Homoscedastic |
| R4-Rep-2 | 0,833 | Homoscedastic |
| R5-Rep-2 | 0,741 | Homoscedastic |
| R6-Rep-3 | 0,359 | Homoscedastic |
| R7-Rep-2 | 0,246 | Homoscedastic |
| R8-Rep-2 | 0,415 | Homoscedastic |
| R9-Rep-2 | 0,665 | Homoscedastic |

Source: Authors (2023).

CONCLUSION

From the results obtained in this work, it can be concluded that GP can obtain mathematical models from data obtained by a DOE experimental array with performance always superior to OLS. Genetic Programming produces mathematical models that are sometimes differentiated when several replicates are performed, but always with similar explanatory power and with biased characteristic that does not affect in any way the quality of prediction of the dependent variable being studied.

REFERENCES

- Alrubaie, K., Godefroid, L., & Lopes, J. (2007). Statistical modeling of fatigue crack growth rate in Inconel alloy 600. *International Journal of Fatigue*, 29(5), 931–940. <https://doi.org/10.1016/j.ijfatigue.2006.07.013>
- Amir Haeri, M., Ebadzadeh, M. M., & Folino, G. (2017). Statistical genetic programming for symbolic regression. *Applied Soft Computing*, 60, 447–469. <https://doi.org/10.1016/j.asoc.2017.06.050>
- Araujo, M. J. F. de, Araújo, M. V. F. de, Araujo Jr, A. H. de, Barros, J. G. M. de, Almeida, M. da G. de, Fonseca, B. B. da, Reis, J. S. D. M., Barbosa, L. C. F. M., Santos, G., & Sampaio, N. A. D. S. (2021). Pollution Credit Certificates Theory: An Analysis on the Quality of Solid Waste Management in Brazil. *Quality Innovation Prosperity*, 25(3), 1–17. <https://doi.org/10.12776/qip.v25i3.1574>
- Beena, V. T., & Kumaran, M. (2010). Measuring inequality and social welfare from any arbitrary distribution. *Brazilian Journal of Probability and Statistics*, 24(1). <https://doi.org/10.1214/08-BJPS022>
- Box, G. E. P., & Wilson, K. B. (1951). On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(1), 1–45. <http://www.jstor.org/stable/2983966>
- Burnham, K. P., & Anderson, D. R. (2004). Model Selection and Multimodel Inference. In K. P. Burnham & D. R. Anderson (Eds.), *Springer New York*. Springer New York. <https://doi.org/10.1007/b97636>

- Ch'ng, C. K., Quah, S. H., & Low, H. C. (2005). A New Approach for Multiple-Response Optimization. *Quality Engineering*, 17(4), 621–626. <https://doi.org/10.1080/08982110500225505>
- Chau, K. (2017). Use of Meta-Heuristic Techniques in Rainfall-Runoff Modelling. *Water*, 9(3), 186. <https://doi.org/10.3390/w9030186>
- Chen, V. C. P., Tsui, K.-L., Barton, R. R., & Meckesheimer, M. (2006). A review on design, modeling and applications of computer experiments. *IIE Transactions*, 38(4), 273–291. <https://doi.org/10.1080/07408170500232495>
- Chen, X., Mei, C., Xu, B., Yu, K., & Huang, X. (2018). Quadratic interpolation based teaching-learning-based optimization for chemical dynamic system optimization. *Knowledge-Based Systems*, 145, 250–263. <https://doi.org/10.1016/j.knosys.2018.01.021>
- Cribari-Neto, F., & Lima, M. da G. A. (2014). New heteroskedasticity-robust standard errors for the linear regression model. *Brazilian Journal of Probability and Statistics*, 28(1). <https://doi.org/10.1214/12-BJPS196>
- Derringer, G., & Suich, R. (1980). Simultaneous Optimization of Several Response Variables. *Journal of Quality Technology*, 12(4), 214–219. <https://doi.org/10.1080/00224065.1980.11980968>
- Diniz, C. A. R., Louzada-Neto, F., & Morita, L. H. M. (2012). The multiplicative heteroscedastic Von Bertalanffy model. *Brazilian Journal of Probability and Statistics*, 26(1). <https://doi.org/10.1214/10-BJPS120>
- Draper, N. R., & Smith, H. (1998). Applied Regression Analysis. In *Wiley-Interscience* (3^a). Wiley-Interscience.
- Dubčáková, R. (2011). Eureka: software review. *Genetic Programming and Evolvable Machines*, 12(2), 173–178. <https://doi.org/10.1007/s10710-010-9124-z>
- Garg, A., & Lam, J. S. L. (2015). Improving environmental sustainability by formulation of generalized power consumption models using an ensemble based multi-gene genetic programming approach. *Journal of Cleaner Production*, 102, 246–263. <https://doi.org/10.1016/j.jclepro.2015.04.068>
- Gomes, F. M., Pereira, F. M., Silva, A. F., & Silva, M. B. (2019). Multiple response optimization: Analysis of genetic programming for symbolic regression and assessment of desirability functions. *Knowledge-Based Systems*, 179, 21–33. <https://doi.org/10.1016/j.knosys.2019.05.002>
- Gonçalves, K. C. M., & Ghosh, M. (2022). Unit level model for small area estimation with count data under square root transformation. *Brazilian Journal of Probability and Statistics*, 36(1). <https://doi.org/10.1214/21-BJPS513>
- Ivanov, D., Dolgui, A., Sokolov, B., Werner, F., & Ivanova, M. (2016). A dynamic model and an algorithm for short-term supply chain scheduling in the smart factory industry 4.0. *International Journal of Production Research*, 54(2), 386–402. <https://doi.org/10.1080/00207543.2014.999958>
- Kalra, M., & Singh, S. (2015). A review of metaheuristic scheduling techniques in cloud computing. *Egyptian Informatics Journal*, 16(3), 275–295. <https://doi.org/10.1016/j.eij.2015.07.001>

Katebi, M., Seif, A., & Faraz, A. (2017). Economic and economic-statistical designs of the T 2 control charts with SVSSI sampling scheme. *Communications in Statistics - Theory and Methods*, 46(20), 10149–10165. <https://doi.org/10.1080/03610926.2016.1231823>

Khuri, A. I., & Conlon, M. (1981). Simultaneous Optimization of Multiple Responses Represented by Polynomial Regression Functions. *Technometrics*, 23(4), 363–375. <https://doi.org/10.1080/00401706.1981.10487681>

Kilmer, J. T., & Rodríguez, R. L. (2017). Ordinary least squares regression is indicated for studies of allometry. *Journal of Evolutionary Biology*, 30(1), 4–12. <https://doi.org/10.1111/jeb.12986>

Lakshmi, K., Mahaboob, B., Rajaiah, M., & Narayana, C. (2021). Ordinary least squares estimation of parameters of linear model. *Journal of Mathematical and Computational Science*. <https://doi.org/10.28919/jmcs/5454>

Liu, G., & Piantadosi, S. (2017). Ridge estimation in generalized linear models and proportional hazards regressions. *Communications in Statistics - Theory and Methods*, 46(23), 11466–11479. <https://doi.org/10.1080/03610926.2016.1267767>

Naderi, M., & Khamenechi, E. (2017). Application of DOE and metaheuristic bat algorithm for well placement and individual well controls optimization. *Journal of Natural Gas Science and Engineering*, 46, 47–58. <https://doi.org/10.1016/j.jngse.2017.07.012>

Pang, S., Zhang, X., & Zhang, Q. (2020). The Hamming distances of saturated asymmetrical orthogonal arrays with strength 2. *Communications in Statistics - Theory and Methods*, 49(16), 3895–3910. <https://doi.org/10.1080/03610926.2019.1591452>

Pinto, E. R., & Pereira, L. A. (2021). On variable selection in joint modeling of mean and dispersion. *Brazilian Journal of Probability and Statistics*, 35(4). <https://doi.org/10.1214/21-BJPS512>

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10), 421–425. [https://doi.org/10.1016/S1364-6613\(02\)01964-2](https://doi.org/10.1016/S1364-6613(02)01964-2)

Poli, R., Langdon, W. B., McPhee, N. F., & Koza, J. R. (2008). A Field Guide to Genetic Programming. In *Lulu Enterprises* (1^o). Lulu Enterprises.

Qasim, M., Amin, M., & Omer, T. (2020). Performance of some new Liu parameters for the linear regression model. *Communications in Statistics - Theory and Methods*, 49(17), 4178–4196. <https://doi.org/10.1080/03610926.2019.1595654>

Salido, M. A., Escamilla, J., Giret, A., & Barber, F. (2016). A genetic algorithm for energy-efficiency in job-shop scheduling. *The International Journal of Advanced Manufacturing Technology*, 85(5–8), 1303–1314. <https://doi.org/10.1007/s00170-015-7987-0>

Salmeron, R., Garcia, C., Garcia, J., & Lopez, M. del M. (2017). The raise estimator estimation, inference, and properties. *Communications in Statistics - Theory and Methods*, 46(13), 6446–6462. <https://doi.org/10.1080/03610926.2015.1125496>

Sampaio, N. A. de S., Reis, J. S. da M., Espuny, M., Cardoso, R. P., Gomes, F. M., Pereira, F. M., Ferreira, L. C., Barbosa, M., Santos, G., & Silva, M. B. (2022). Contributions to the future of metaheuristics in the contours of scientific development. *Gestão & Produção*, 29(1), 1–19. <https://doi.org/10.1590/1806-9649-2022v29e099>

Sathiya, P., Abdul Jaleel, M. Y., Katherasan, D., & Shanmugarajan, B. (2011). Optimization of

laser butt welding parameters with multiple performance characteristics. *Optics & Laser Technology*, 43(3), 660–673. <https://doi.org/10.1016/j.optlastec.2010.09.007>

Silva, A. R. S., Azevedo, C. L. N., Bazán, J. L., & Nobre, J. S. (2021). Bayesian inference for zero-and/or-one augmented beta rectangular regression models. *Brazilian Journal of Probability and Statistics*, 35(4). <https://doi.org/10.1214/21-BJPS505>

Silva, H. de O. G. da, Costa, M. C. M., Aguilera, M. V. C., Almeida, M. da G. D. de, Fonseca, B. B. da, Reis, J. S. da M., Barbosa, L. C. F. M., Santos, G., & Sampaio, N. A. de S. (2021). Improved Vehicle Painting Process Using Statistical Process Control Tools in an Automobile Industry. *International Journal for Quality Research*, 15(4), 1251–1268. <https://doi.org/10.24874/IJQR15.04-14>

Wan, Q., Wu, Y., Zhou, W., & Chen, X. (2018). Economic design of an integrated adaptive synthetic chart and maintenance management system. *Communications in Statistics - Theory and Methods*, 47(11), 2625–2642. <https://doi.org/10.1080/03610926.2016.1271425>

Wang, X., & Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3), 589–611. <https://doi.org/10.1111/rssb.12127>

Wilcox, R. R. (2007). On Flexible Tests of Independence and homoscedasticity. *Journal of Modern Applied Statistical Methods*, 6(1), 30–35. <https://doi.org/10.22237/jmasm/1177992240>

Will M. Bertrand, J., & Fransoo, J. C. (2002). Operations management research methodologies using quantitative modeling. *International Journal of Operations & Production Management*, 22(2), 241–264. <https://doi.org/10.1108/01443570210414338>

Wu, X., & Yang, Z. (2013). Nonlinear speech coding model based on genetic programming. *Applied Soft Computing*, 13(7), 3314–3323. <https://doi.org/10.1016/j.asoc.2013.02.008>

Zhang, J., Gai, Y., Cui, X., & Li, G. (2020). Measuring symmetry and asymmetry of multiplicative distortion measurement errors data. *Brazilian Journal of Probability and Statistics*, 34(2). <https://doi.org/10.1214/19-BJPS432>