

## Sizing an Artificial Neural Network for Modelling Piping Assembly Labor Productivity Under Different Conditions

Marcus Sarmiento, M.Sc. \*, Miguel Luiz Ribeiro Ferreira, D.Sc.\*\*

\*(Civil Engineering Department, Universidade Federal Fluminense, Niterói, Brazil

\*\* (Graduate Program of Industrial Construction, Universidade Federal Fluminense, Niterói, Brazil

### ABSTRACT

Executing projects according to its estimated cost and schedule is fundamental to guarantee the economic and financial viability of the businesses associated with them. In this context, the estimated duration of an activity is key for the development of the project schedule and budget. However, in various occasions, due to the factors affecting Labor Productivity (LP), the data used as input for the estimates is often collected in conditions different than the execution conditions, thus causing significant variations between the Basic Labor Productivity Rate (BLPR) and the Real Labor Productivity Rate (RLPR) of a project. The research contributes to engineering management body of knowledge by developing a model, through the simulation by Artificial Neural Networks (ANN), capable of adjust the BLPR, considering the factors affecting LP present in a certain project. Furthermore, different ANNs configurations were tested and compared, in order to identify the optimal configuration and how each parameter of the ANN affects the result calculated by the model. Given that, future researches will have a start point to size ANN eliminating the need of test and evaluate several scenarios. The research has identified that RLPR was about 16.86% lower than BLPR, while the model output was 2.21% lower than RLPR. The Mean Square Error (MSE) obtained as a result of the simulation was 9.67E-04 and the correlation was 0.91855. Finally, the results showed that several ANN architectures and configurations can be used and, even so, present satisfactory results.

**Keywords** - Artificial Neural Networks (ANN), Influencing factors, Labor productivity, productivity in construction.

Date of Submission: 13-02-2021

Date of Acceptance: 27-02-2021

### I. INTRODUCTION

An adequate estimate of an activity duration is key for a project scheduling as well as the cost estimating [1]. The elaboration of estimates and project planning is based on the experts' opinion, analogous estimates, parametric estimates or a combination of these [2].

Due to factors affecting LP, the data used as input for analogous estimates, as well as for parametric estimates, is often obtained under different conditions than the ones of the project to be estimated [3]. These estimates consider the BLPR, calculated through the average LP obtained in the projects previously performed. Thereby, it is necessary that the BLPR are adjusted to reflect the expected productivity for a given project [4]. This adjustment is made, most of the time, based on the opinion of experts, through the multiplication of the BLPR by adjust factors. It is noted that this decision-making process is highly subjective, requires substantial knowledge and skill of the estimator and has the potential to cause significant fluctuations in schedule and cost.

In order to attenuate this problem, several modeling techniques have been applied [5]. These models can focus on the analysis of activities or on the factors affecting LP [6]. Supervised methods in ANN have significant academic contribution in modeling with a focus on contributing factors [7].

In this context, this research developed a model, through the use of ANN, capable of adjusting the BLPR for piping assembly taking into account the influencing factors. Furthermore, one hundred and eight different architectures of feedforward back-propagation network (FFNN) were tested and compared, in order to identify the impact of each variable of the ANN architecture on the result obtained by the model. Specifically, the following variables were tested: number of layers, number of neurons, activation and training functions.

### II. ARTIFICIAL NEURAL NETWORKS (ANN)

ANNs are information processing systems and its performance characteristics resemble biological neural networks. An ANN essentially consists of several artificial neurons organized in layers [8]. The ability to learn of ANNs is one of the

most important qualities of these structures and is defined as the ability of ANNs to adapt, according to pre-existing rules, changing its performance over time [9]. The designer of an ANN should take care of the definition of network architecture and the appropriate training algorithm to solve given problem [11]. FFNN have been used successfully in several recent studies related to productivity in construction [2,3,12] The typical structure of an FFNN can be seen in Fig.1.

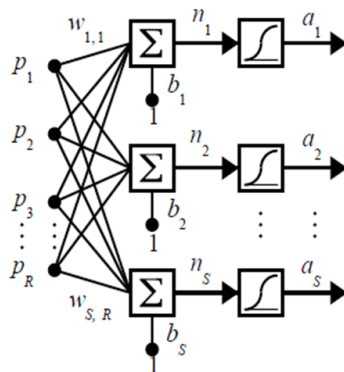


Figure 1 – Single layer FFNN

In a FFNN the input neurons will compute a given input weighted by the synaptic weight and add the result to the neuron bias [11]. Finally, the neuron will use the activation function to send the result to the neuron in the next layer. The result obtained in the output layer will be compared with the data used as a training reference. This operation will be repeated cyclically until one or more stop criteria, previously established, is reached.

### III. RESEARCH METHODOLOGY

Given the objective of developing a model capable of adjusting BLPR taking into account the factors that affect LP, this research was divided into 4 phases: (a) identification of factors affecting LP, (b) data collection and normalization, (c) architecture development and (d) ANN training and validation.

#### IDENTIFICATION OF FACTORS AFFECTING LP

The identification of contributing factors was carried out based on a wide bibliographic review and field research [13]. As a result of his research, Sarmiento identified 56 factors that contribute to the performance of construction productivity in Brazil. These factors were listed according to their relevance. The importance index [14] was used as a ranking criterion. Table 1 lists the 14 main contributing factors identified by the author, with respect to the piping assembly, which were used as basis to compose the survey questionnaire.

Table 1 – Contributing factors

#	Contributing Factors
1	Material availability (pipe spools)
2	Material availability (valves and fittings)
3	Material preparation and deliver
4	Project complexity
5	Planning
6	Tool availability
7	Welding consumables
8	Quality of raw material
9	QA/QC availability
10	Team experience
11	Quality requirement level
12	Overtime
13	Lack of support equipment
14	Rework
15	Others

#### DATA COLLECTION AND NORMALIZATION

Data collection was carried out in a construction site located in Minas Gerais, Brazil. Data collection goal was: (a) identify the RLPR during construction and (b) identify factors affecting LP. Productivity concept was based on the partial productivity factor, described as direct LP [14]. The LP (1) measures the productivity of specific activities by monitoring field activities and is usually expressed as the number of outputs (Q) over the amount of labor (L).

$$LP = \frac{Q(\text{production units})}{L(HH)} \quad (1)$$

This factor is in line with the BLPR adopted by the estimating department, when bidding the project. Direct labor is all the labor required to perform services, up to the foreman level [16]. Finally, the concept of global productivity [17] is considered in the productivity survey phase. The RLPR surveyed in the field was compared with the BLPR. As a result of this comparison, a productivity factor was calculated and used as an input to the model.

The identification of factors affecting LP was carried out through the daily filling of a form composed of 14 closed and one open question, where the factors of greatest importance index identified in the bibliographic research were listed. The questionnaires were completed by the construction supervisors. In this questionnaire, the respondent was invited to grade each factor on a scale of 5 to -5. The respondent also answered whether the influence is positive or negative. After the questionnaires are applied, the answers must be

normalized [12]. This normalization is necessary so that linguistic factors are converted into numerical factors between 1 and -1. Given the tool adopted for the elaboration of the model, data related to productivity will be expressed in decimal numerical form. Table 2 shows the conversion of linguistic terms to numeric scale and standardized numerical scale successively.

### DEVELOPMENT OF ANN ARCHITECTURE

The MATLAB 2019b Deep Learning toolbox was used to create, train, validate and test the ANN. The choice of this tool comes from the quality and variety of the output reports available, as well as from the wide range of formatting options of ANN. Added to this is the low cost of obtaining the software and the high degree of use in the academy. Considering the variables mentioned by Beale et al. [11] this research tested 108 different FFNN configurations. The different scenarios were defined in order to explore the widest possible range of combinations of: (a) number of neurons per layer, (b) training function and (c) activation functions.

Table 2 - Example of data normalization

Types of factors	Linguistic description	Num. Scale	Stand. Scale
Positive factors	Very strong effect	5	1.0
	Strong effect	4	0.8
	Moderate effect	3	0.6
	Little effect	2	0.4
	Negligible effect	1	0.2
Negative factors	Negligible effect	-1	-0.2
	Little effect	-2	-0.4
	Moderate effect	-3	-0.6
	Strong effect	-4	-0.8
	Very strong effect	-5	-1.0

The scenarios related to the number of hidden layers neurons were tested according to the Kolmogorov and Fletcher-Gloss theorems. According to the Kolmogorov method, the number of neurons is expressed below (2). Therefore, the number of neurons  $n$  will be equal to 31.

$$n = 2n1 + 1 \quad (2)$$

When calculated using the Fletcher-Gloss method (3) the number of neurons ( $n$ ) will be equal to 9.

$$2\sqrt{n1} + 1 \leq n \leq 2n2 + 1 \quad (3)$$

Regarding the number of hidden layers, one and two layers ANNs were tested [10].

Finally, the tangent sigmoid, sigmoid and linear activation functions were tested as proposed by Heravi and Eslamdoost [12].

### ANN TRAINING AND VALIDATION

As it is a neural network with supervised learning, in the training of a FFNN, each training example will be accompanied by a desired value. In the specific case of this research, the desired value is the productivity factor described in the data collection and normalization section.

Being  $t_j(t)$  the desired response to neuron  $j$  at time  $t$  and  $a_j(t)$  the observed response of neuron  $j$  at time  $t$ , obtained through a stimulus  $x(t)$  present at the input of the neural network. Being  $e(t) = t_j(t) - a_j(t)$  the error found at the output of neuron  $j$  at time  $t$ , the supervised learning process seeks to make it as close to zero as possible. For this, the algorithm develops a comparative map between inputs and outputs by minimizing the MSE (4) where  $n$  is the number of samples,  $ti$  is the model's output related to the  $n$ th output and  $ai$  is the desired output.

$$EQM = \frac{\sum_{i=1}^n (ti - ai)^2}{n} \quad (4)$$

During the training process, the correlation between the model's outputs and the expected output should also be observed [3]. Correlation is defined as the measure of the direction and degree with which two variables,  $X$  and  $Y$ , are linearly associated in a sample. Mathematically, the correlation measure is expressed below (5).

$$R = \frac{\sum_{i=1}^n XiYi - (\sum_{i=1}^n Xi)(\sum_{i=1}^n Yi)}{\left\{ \left[ n \sum_{i=1}^n Xi^2 - (\sum_{i=1}^n Xi)^2 \right] \cdot \left[ n \sum_{i=1}^n Yi^2 - (\sum_{i=1}^n Yi)^2 \right] \right\}^{1/2}} \quad (5)$$

It is important to note that during network training, where the supervised training paradigm is used, generalization errors can generally be found, the most common among them being over adjustment, which is characterized by very low level of error in the training phase and high errors in the testing phase, or during operation [12]. The network memorizes the responses to the training cases but is unable to generalize when submitted to new cases.

In order to avoid generalization errors, Bayesian regularization and early stop techniques will be tested. Bayesian normalization presented by MacKay [18] is used as a reference by the *trainbr* function available on MATLAB 2019b. Bayesian networks are diagrams that organize knowledge in a given area through a mapping between causes and effects.

The criteria of early stop aims to determine, forcibly, the appropriate number of epochs or times for a given ANN. In a certain point in the training of multilayered ANN, the MSE starts a growth process, after reaching its minimum value [11]. The MATLAB 2019b *trainscg* function avoids this problem by stopping the training process early, in other words, when the system observes that the MSE

grows for a certain consecutive number of epochs, the process is interrupted and the synaptic weight and bias values identified as minimum are returned for ANN. When the *trainscg* function was used, 70% of the data was used in ANN training, 15% for validation and 15% for testing. Since the *trainbr* function does not require testing, the samples were divided between training and validation. Following the 75/25 percentage [7].

#### IV. INFLUENCE OF ANN PARAMETERS AND SCENARIO SELECTION

Through the analysis of overall results, it can be observed that ANNs are a suitable modeling tool for adjusting the LP curves. However, the parameters adopted to measure the ANN performance (MSE and correlation) are not enough to identify the best configuration. In this context, to recognize the best ANN configuration, it is necessary to analyze the absolute errors, found when a specific parameter is adopted. These errors were plotted as box and whisker chart, which shows distribution of data into quartiles, highlighting the mean and outliers. In this sense, we seek to identify the data set with the lowest interquartile range, the lowest upper limit and the higher lower limit.

Absolute errors were calculated for each of the 84 simulations performed in the field. Once the ANN is trained, the software MATLAB 2019b generates a table of results for each simulation. This table of results is compared to RLPR. The difference found is the absolute error. The next step is the segregation of the parameter to be evaluated. The process is repeated for all parameters analyzed.

The activation functions tried at input layer, hidden layer and output were linear, sigmoid and tangent sigmoid. When the MSE was evaluated it was observed that there was not significant variation caused by the adopted function. Also, the correlation found was consistent for all functions at all layers. The charts plotted at Fig. 2, 3 and 4 shows that, although the average and median, present close values the linear function presented less dispersion of absolute errors. In this sense, given the equivalence found in the other parameters, the linear function will be adopted as an activation function for all layer.

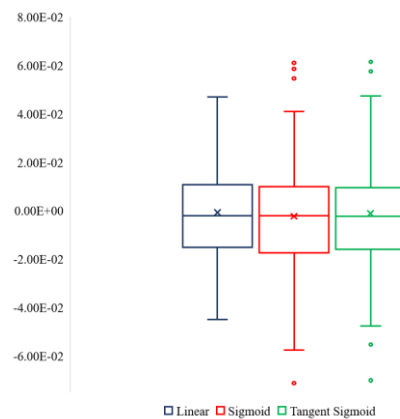


Figure 2 – Box and whisker chart input layer

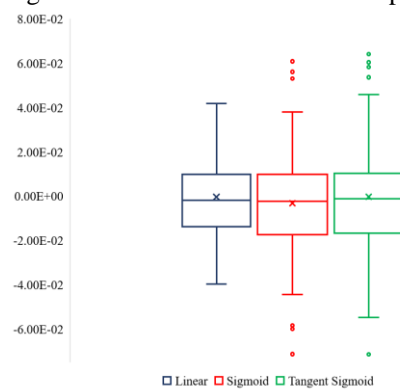


Figure 3 – Box and whisker chart hidden layer

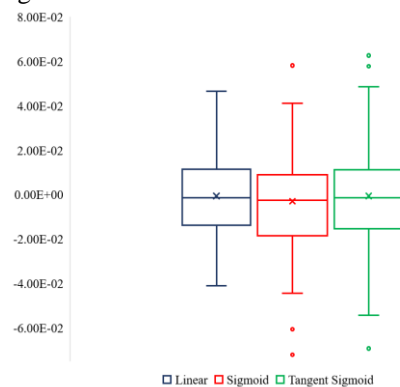


Figure 4 – Box and whisker chart output layer

With regards to number of neurons scenarios with 9, 15 and 31 neurons were tested. As identified for the activation function MSE and correlation does not play prominent role.

Given that, when the absolute error is evaluated the scenarios where 31 neurons were used show the best result as can be seen on fig. 5.

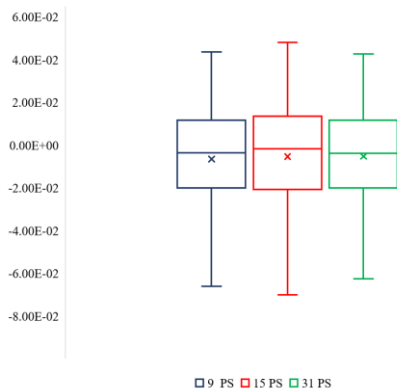


Figure 5 – Box and whisker chart number of neurons

The last parameter to be evaluated was the training function. When compared the MSE correlation and even the dispersion of absolute errors the results were extremely similar. So, the criteria number of epochs was considered to define the training function to be used. In this context, the average of epochs found when the function *trainbr* was used was 266 while the average for *trainscg* function was 18. Considering that a higher number of epochs requires more computational capacity the function *trainscg* was adopted.

## V. RESULT ANALYSIS

### ANN PERFORMANCE

Once that the parameters are defined the scenario number B200 was chosen and its performance was evaluated. Specifically, the correlation and MSE. Given that, the *trainsig* was adopted as training function the early stop technic is utilized to determine when the ANN training is complete, and its result is optimal. As observed at Fig. 6 the best validation performance was achieved at epoch 53 and the early stop was at epoch 59. The MSE obtained for the overall simulation was  $9.67E-04$  while the best performance during the trained was  $6.84E-04$ . Moreover, it can be observed that training, validation and test curves have similar shapes, which denotes the consistency between the samples chosen by the software to each curve.

The overall correlation is 0.91855. which indicates that 91.855% of the samples satisfies the function during the overall evaluation. Correlation found during training, validation and test were 0.9555, 0.89684 and 0.65091 respectively.

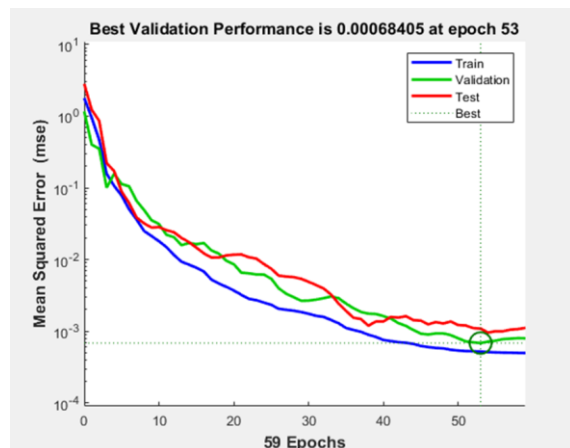


Figure 6 – Training, validation and test performance

The correlation results reveal a good adherence between the model and the data collected on site, RLPR. Fig. 7 shows the regression and correlation plotted by the software MATLAB 2019b.

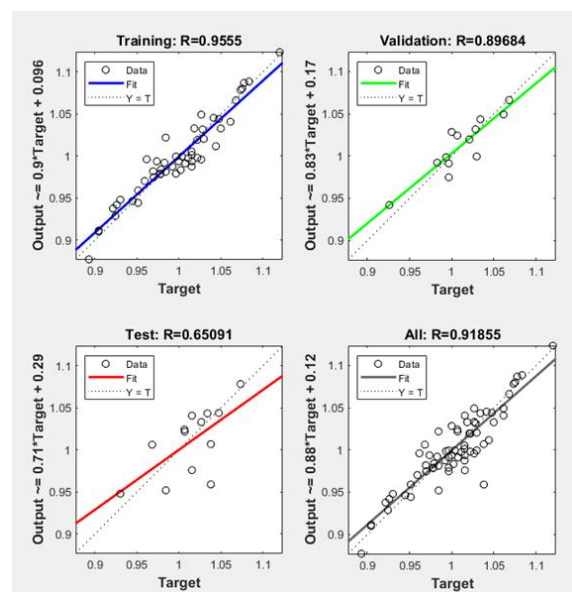


Figure 7– Regression and correlation

Based on the results presented previously the performance of training, validation and test of the ANN is consistent with previous research [2,3,7] as well as it's a good fit for the propose of the research.

### EVALUATION OF MODEL OUTPUT

Once the ideal scenario is identified and its performance is evaluated the maximum, minimal and adjusted LP curves were generated by the ANN. Fig. 7 represent the generated curves as well as the BLPR and RLPR.



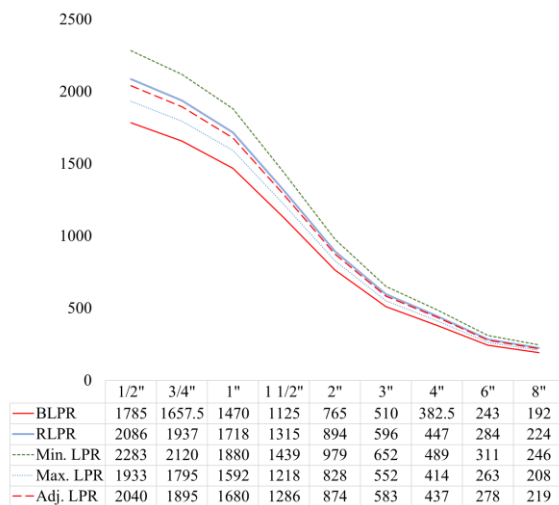


Figure 7 – Labor productivity curves

BLPR is the LP estimated at the time of the project bid, it's the expected productivity. The RLPR is the productivity measured during the construction phase. Max. and Min. LPR are calculated by the ANN taking in account the samples used during the training process. The Adj. LPR is the result generated by the model.

When compared the Adj. LPR and RLPR the variation found was 2.21%. its important to notice that the difference between BLPR and RLPR is 16,86% which mean that the model was able to reduce 89,92% the losses related to factors affecting LP.

Given that the Max. LPR and Min. LPR range is 9,60% its necessary to understand Cumulative Distribution Function (CDF) for the results calculated by the ANN. Again, the results from the 84 simulations will be used as basis. In this sense, the software @risk 8.1.1 was selected to support this activity. The Fig.8 plots the CDF and shows that there is a 90.0% probability of the absolute error calculated by the model be allocated between - 4,19% and 4,17%. Furthermore, the probability of the error calculated by the model to be zero is greater than 50%.

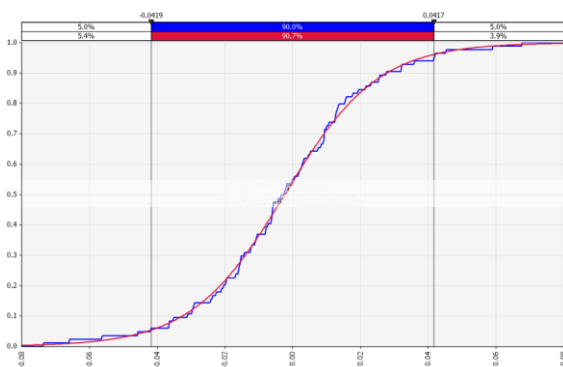


Figure 7– CDF scenario error scenario B200

## VI. CONCLUSION

An ANN was sized, and the impact of its parameters were evaluated. As result it was identified that the tested parameters will not impact significantly the performance or result calculated by the ANN. However, absolute error dispersion as well as the epochs necessary to reach the stopping criteria can be affected.

Through the analysis of the results presented, it is concluded that the ANN is an adequate tool for the elaboration of models that will adjust BLPR considering the factors affecting LP. Furthermore, it was observed that the model was able to predict the RLPR within a  $\pm 4,2\%$  error. The MSE identified for the adopted scenario is  $9.67E-04$  though the correlation is 9.1855.

This is the first of a kind research considering adjusting the BLPR for piping assembly. The construction industry can use the trained ANN as a tool for cost and schedule development. Other academic research may take the advantage of the extensive research to comprehend how each parameter affected the result of the ANN.

The downside of the method is that the ANN learns from examples, which means that some calibration may be needed to adapt the ANN to other companies BLPR. Once its done the generalization will be achieved.

## REFERENCES

- [1]. Project Management Institute. (2004). *A guide to the project management body of knowledge (PMBOK Guide) (Third ed.)*. Newton Square, PA: Project Management Institute
- [2]. El-Gohary, K. M., Aziz, R. F., and Del-Khalek, H. A. (2017). "Engineering Approach Using ANN to Improve e Predict Construction Labor Productivity under Different Influences", *Journal of Construction Engineering e Management*, 143(8).
- [3]. Dayanand, A., and Shanmugapriya, S. (2016). "ANN Model for estimation of construction labor productivity", *International Journal of Modern Trends in Engineering and Science*, 3(8), 9.
- [4]. Mani, N., Kisi, K. P., Rojas, E. M., and Foster, T. E. (2017). "Estimating Construction Labor Productivity Frontier: Pilot Study", *Journal of Construction Engineering and Management*, 143(10).
- [5]. Yi, W., and Chan, A. P. C. (2014, Mar). "Critical Review of Labor Productivity Research in Construction Journals", *Journal of Management in Engineering*, 30(2), 214-225.

- [6]. Tsehayae, A. A. and Fayek (2016). System model for analysing construction labour productivity. *Construction Innovation* 25(03).
- [7]. Gerek, İ. H., Erdis, E., Mistikoglu, G., and Usmen, M. (2014). “Modelling masonry crew productivity using two artificial neural network techniques”, *Journal of Civil engineering and Management*, 21(2), 231-238.
- [8]. Kriesel, D. (2007). “A Brief Introduction to Neural Networks”.
- [9]. Rauber, T. W. (1997). “Redes Neurais Artificiais”, Universidade Federal do Espírito Santo, Brazil
- [10]. Hagan, M. T., Demuth, H. B., Beale, M. H., and De Jesus, O. (2014). *Neural Network Design*, 2nd Edition.
- [11]. Beale, M. H., Hagan, M. T., and Demuth, H. B. (2010). *Neural Network Toolbox™ 7*, 951.
- [12]. Heravi, G., and Eslamdoost, E. (2015). “Applying Artificial Neural Networks for Measuring and Predicting Construction-Labor Productivity”, *Journal of Construction Engineering and Management*, 141(10).
- [13]. Sarmiento, M. V. (2015). “Fatores contribuintes para o desempenho da produtividade em construção e montagem em projetos industriais no Brasil”, Master’s dissertation, Universidade Federal Fluminense, Rio de Janeiro, Brazil.
- [14]. Lim, E., and Alum, J. (1995). “Construction productivity: issues encountered by contractors in Singapore”, *International Journal of Project Management*, 13(1), 51-58.
- [15]. Jarkas, A. M., and Bitar, C. G. (2012). “Factors affecting construction labor productivity in Kuwait”, *Journal of Construction Engineering and Management*, 138(7), 811-820.
- [16]. Ferreira, M. L. R; Morano, C. A. R; Freire, I. F. (2010). PROMINP (Programa de Mobilização da Indústria do Petróleo e Gás Natural). Projeto E&P 27.5 – *Métricas de Desempenho da Indústria Nacional. Padrões de Métricas da Indústria EPC Nacional*.
- [17]. Tabim, P. M. (2013). “*Estimativa de Produtividade na Soldagem de Dutos Terrestres com Aplicação de Simulação de Monte Carlo*”, Master’s dissertation, Universidade Federal Fluminense, Rio de Janeiro, Brazil.
- [18]. Mackay, D. J. C. (1992, May). “Bayesian Interpolation”, *Neural Computation*, 4(3), 415-447.