PARAMETRIC COST MODELLING OF COMPONENTS FOR TURBOMACHINES: PRELIMINARY STUDY

Campi, Federico (1);
Mandolini, Marco (1);
Santucci, Federica (1);
Favi, Claudio (2);
Germani, Michele (1)

1: Università Politecnica delle Marche;
2: Università degli studi di Parma

ABSTRACT

The ever-increasing competitiveness, due to the market globalisation, has forced the industries to modify their design and production strategies. Hence, it is crucial to estimate and optimise costs as early as possible since any following changes will negatively impact the redesign effort and lead time.
This paper aims to compare different parametric cost estimation methods that can be used for analysing mechanical components. The current work presents a cost estimation methodology which uses non-historical data for the database population. The database is settled using should cost data obtained from analytical cost models implemented in a cost estimation software. Then, the paper compares different parametric cost modelling techniques (artificial neural networks, deep learning, random forest and linear regression) to define the best one for industrial components.
Such methods have been tested on 9 axial compressor discs, different in dimensions. Then, by considering other materials and batch sizes, it was possible to reach a training dataset of 90 records. From the analysis carried out in this work, it is possible to conclude that the machine learning techniques are a valid alternative to the traditional linear regression ones.

Keywords: Design costing, Machine learning, Early design phases, Conceptual design

Contact:
Campi, Federico
Università Politecnica delle Marche
DIISM
Italy
f.campi@pm.univpm.it

Cite this article: Campi, F., Mandolini, M., Santucci, F., Favi, C., Germani, M. (2021) 'Parametric Cost Modelling of Components for Turbomachines: Preliminary Study', in *Proceedings of the International Conference on Engineering Design (ICED21),* Gothenburg, Sweden, 16-20 August 2021. DOI:10.1017/pds.2021.499

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press

1 INTRODUCTION AND LITERATURE REVIEW

The product conceptualisation is the first phase of a design process, which then continues with the embodiment and detailed design. During this phase, designers collaborate to define the overall product architecture and individual modules or components' general characteristics. The definition of a product architecture starts from modules arrangement and layout definition, including the analysis of the preliminary manufacturing process of each element (Ulrich et al., 2011). To facilitate the design process, 2D CAD models are often developed for the graphical representation of components. Cost reduction opportunities are significant at the conceptual stage because there is enough space to investigate different alternatives (product architectures). More than 70% of the product cost is committed during the conceptual design stage (Boothroyd et al., 2011). Hence, it is crucial to estimate and optimise costs as early as possible since any following changes will negatively impact the redesign effort and lead time (Favi et al., 2016). During the conceptual design, cost estimation cannot be performed using analytical cost estimation approaches based on 3D CAD models as they are not yet available. On the contrary, if such models are available, they do not contain the details required to get reliable economic results with such methods (Mandolini et al., 2020). The evaluation needs to follow other approaches capable of processing 2D geometries or simple numerical parameters. This paper aims to compare different parametric cost estimation methods that can be used for analysing mechanical components. The methods developed for cost estimation are grouped into two families: (i) qualitative methods, which include knowledge-based and intuitive methods, and (ii) quantitative methods, which could be divided into analytical methods and parametric methods (Niazi et al., 2005). In the context of parametric methods, the scientific literature is characterised by several scientific papers aiming at applying and evaluating the performance of data mining, machine learning and artificial intelligence approaches for cost estimation during the preliminary design phases. Regression (linear or not) is one of the most widespread methods for parametric cost estimation. Regression models can learn from the given data by adjusting the regression parameters to map a mathematical relationship based on the given data. This method attempts to establish the nature of the relationship between variables by providing a prediction mechanism. There are two different types of variables: dependent, which represent various system parameters, and independent, which represent the costs of the project or the part. Regression, therefore, is a branch of applied statistics that allows to quantify the relationship between the dependent variable and one or more independent variables and to describe the accuracy of this relationship. Within the scientific literature, several applications were developed for parametric cost modelling using linear regression. Langmaak et al. (Langmaak et al., 2013) present a cost estimation tool for gas turbine components. The tool is divided into two parts. The first is a generic factory cost model based on activity-based costing that can estimate various costs at multiple levels of any manufacturing plant. A parametric and scalable cost model is the second tool for assessing the unit cost of future integrally bladed disc (blisk), a component used by the aerospace industry in gas turbine compressors. Another example in the context of gas turbine engineering is presented by Masel et al. (Masel et al., 2010), with a definition of a cost estimation model based on CER (Cost Estimation Relationship. Other examples of parametric cost estimations for jet components are provided by Bertoni et al. (Bertoni et al., 2018; Bertoni et al., 2020). Parametric cost estimation using regression models is also applied in machining (Stockton et al., 2013) in software development (Heiat, 2002), in sheet metal parts (Verlinden et al., 2007), or painting cost estimation (Stockton et al., 2013). In parametric cost estimation also machine learning (ML) techniques play an essential role. Machine learning is generally more efficient than traditional mathematical and statistical models in manufacturing. However, enterprises are still hesitant in adopting these techniques because they have the limit of being considered as "black boxes" [Hihn, 2015]. It is not possible to give a theoretical interpretation of the results, especially in unexpected or unjustified values. On the other hand, a linear regression model can be deducible from technical considerations and, consequently, it is clearly understandable for the users. Traditional methods remain incapable of understanding complex relations among data samples' features and predicting unknown feature values for a new piece (Dogan et al., 2011). ML techniques include artificial neural network (ANN), deep learning (DL), support vector machines (SVM), decision trees (DT) and random forest (RF). ANN are biologically inspired models to mimic the human neural system for information-processing and computation purposes. ANN is a ML technique that can learn from past data. Learning forms can be supervised, unsupervised, and reinforcement learning (Sala et al., 2018). ANN is widely used in many fields, as

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press

cost estimation of construction projects (Elmousalami, 2019), in software development effort estimation (Heiat et al., 2002) and also in cost estimation of an industrial component or part (sheet metal parts (Verlinden et al., 2007), injection moulding (Wang et al., 2013), aircraft (Chen et al., 2020) and machining (Ning et al., 2020)). Some authors compared the results of different parametric cost estimation methods, in particular, ANN. Although such techniques are more challenging to interpret (Loyer et al., 2016), they generally obtain better results (mean square error - MSE and mean absolute percentage error - MAPE) than CER (Cavalieri et al., 2004). Regression-based methods require the definition of a relationship between inputs and outputs, while in ANN, this relationship happens automatically (Cavalieri et al., 2004). If the number of inputs is limited, the regression works well. On the contrary, with many inputs, the ANN is the most suitable choice (Verlinden et al., 2007). Another popular ML algorithm is the Random Forest, known for its simplicity, ease of use and interpretability. Random forests model is an ensemble method of decision trees, a weak learner and can easily be over-fitting. By assembling the decision trees, their instability and high variance can be overcome (Wang et al., 2018). A decision tree represents a classification or regression model in a tree structure. Each node in the tree structure represents a particular "question" about a feature; each branch signifies a decision. At the end of a branch, each leaf is the corresponding output value (Breiman, 2001). To obtain a result, starting from a specific input, the decision process begins from the root node (at the top) and runs through the tree until it reaches a leaf that contains the result. In each node, the path to follow depends on the values assumed by the various features. Similar to neural networks, the tree is created through a learning process using training data. RF algorithms are used for cost estimation in the construction industry (Bilal et al., 2020), in software estimation effort (Abdelali et al., 2019), in the marine field (Isıklı et al., 2020) or for battery capacity estimation (Li et al., 2018). However, there is a lack of RF application for cost estimation of industrial components. Generally, ML algorithms and regression methods are trained using historical data. Then they use present data to predict future outcomes. However, it must be considered that data acquisition can be changed and adapted over time. Therefore, it may be necessary to clean up historical data or models based on older data must be retrained (Weichert et al., 2019). Historical data may be few, with a not well-defined structure. They may contain outliers, so their use is not always applicable.

Based on the above literature analysis, the first research question could then be summarized in: "*Is it possible to overcome the historical data limitations?*". In response to this research question, the current work presents a cost estimation methodology that uses non-historical data for the database population. The database is populated using should cost data obtained from an analytical cost model implemented in a cost estimation software (LeanCOST® by HyperLean).

The second research question is: "*What's the best parametric cost estimation approach for industrial components?*" Concerning this latter, the article is focused on a comparison between machine learning techniques (ANN, DL and RF) and regression one (CER). The goal is to define the best performing cost estimation method for a well-defined family of industrial components: axial compressor discs.

2 MATERIALS AND METHODS

This paper's research work is ground on a methodology (Figure 1) adopted to perform the parametric cost estimation.

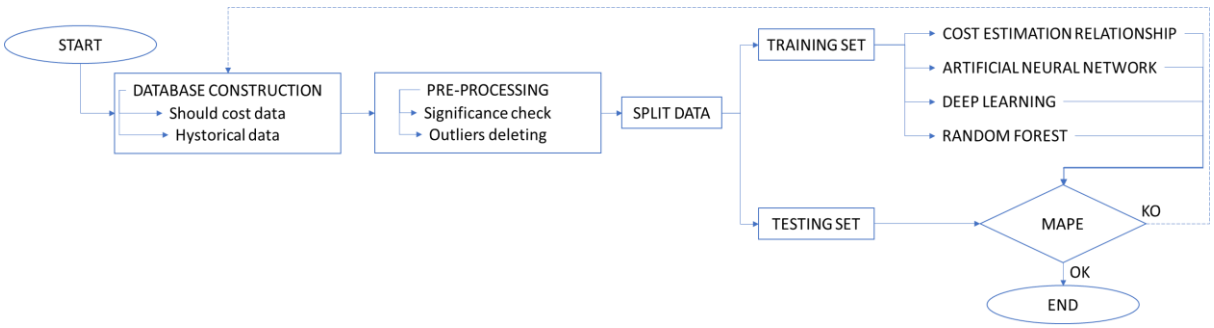


Figure 1: Parametric cost estimation methodology

The first step of this methodology consists of defining and pre-processing the database (Section 2.1). The dependent parameter (the cost of the parts) came from a should-cost analysis. This process aims to determine the price of a piece considering factors such as the raw materials cost, production costs,

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press

overhead, etc. The usefulness of this type of analysis is twofold. On the one hand, there is the "contracting with the supplier" factor. On the other hand, when the part is produced internally, the analysis allows identifying the most burdensome operations in terms of cost. Starting from a set of should cost data, a pre-processing phase allows preparing the database used as a training dataset. In particular, any outliers are excluded, and only cost drivers with a strong influence on the output are considered. A significance check analysis defines the main cost drivers. Subsequently, the database is split into two groups: the training and testing sets. The training set is used to implement the traditional linear regression method (Section 2.2) and non-linear machine learning methods (Sections 2.3 and 2.4). The testing set is used to validate the resulting cost models (Section 2.5). The final validation is based on the value of the prediction error.

2.1 Database construction and pre-processing

The accuracy of a cost estimate analysis depends highly on the form of the input database. The database consists of input data (geometric and non-geometric drivers) and related output data (raw material and production costs). The first step is database generation. The database is populated using should cost data obtained from analytical cost models implemented in a cost estimation software (LeanCOST® by Hyperlean). Once the database is created, the pre-processing step begins. This phase involves two main activities:

- Check the significance of each input variable.
- Identify and delete outliers.

The way to check each input's significance occurs through a regression model and the Pearson coefficient definition to evaluate the correlation between variables. At first, input variables that are not correlated with outputs (Pearson values < 0.3) are identified and removed. After that, input variables positively related to other input ones (Pearson values > 0.8) are further defined. Such variables can be deleted because they do not represent a cost-driver or output (Duran et al., 2012). Once finalised the significance check, outliers must be defined. Outliers correspond to wrong and also duplicate and incomplete records. To obtain an efficient cost estimation, it is necessary to delete all outliers. Cook's distance is one of the methods for outlier's discovery. If Cook's values are < 1, there is no need to delete that case; otherwise, the outliers must be removed (Cook et al., 1982). The next step consists of splitting the data. Data are divided into two groups: a part is used to create the cost model, while the rest is used for its testing. The test set allows validating the model results. A typical subdivision is 60% for training, and 40% for testing (Green et al., 1991).

2.2 Cost estimating relationship (CER)

The first parametric approach implemented in this work is based on linear regression. Since this technique can only be developed with a numerical database, the first step was to convert all the categorical variables (contained in the starting database) into a dummy or binary variables. The use of dummy variables, indicated by D, is the standard method to solve qualitative factors' value assignment. The D value definition could be a manual process, using "0" or "1" according to its dichotomy characteristics. D=1 means that the qualitative factors possess specific attributes or are subjected to some aspects. D=0 is the contrary (Tan et al., 2011). Once the nature of starting variables has been changed, the procedure developed is iterative. Through regression, the parametric equation (CER) is determined and validated. The parameter that allows measuring CER accuracy is the coefficient of determination R2. If its value is acceptable, the analysis shall be completed. Otherwise, the procedure shall be repeated after making appropriate changes to the database. R2 values higher than 0.9 mean a satisfying correlation. R2 values between 0.6 and 0.9 should push analysts to deep dive into the data to identify more cost drivers. R2 values lower than 0.6 are not acceptable. No correlation can be used for further studies, so the analysis must be repeated (Martinelli et al., 2019).

2.3 Artificial Neural Networks (ANN) and Deep Learning (DL)

Besides the CER method, this work considers two other approaches: the artificial neural network (ANN) and deep learning (DL). While the CER development requires conversion to dummy variables, ANN and DL can handle numerical and categorical variables. To implement these ML techniques, it is necessary to define two significant parameters: the number of hidden layers and the number of neurons for each hidden layer. On this aim, the theory of Haytham H. Elmousalami (Elmousalami et al., 2019)

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press

has been applied. According to this theory, one or two hidden layers are most likely to converge. Too more or too less may lead to poor convergence results. Empirically speaking, one layer may be chosen for the general problems, and two layers may be used for more complex ones. The definition of the right quantity of neurons for the hidden layer is also crucial. A low number of neurons will reduce the resources needed to solve the problem. Using too many neurons, the training effort will significantly increase. Besides, an excessive number of hidden neurons may cause a problem called overfitting. One rough guideline for choosing the right number of hidden neurons in many problems is the geometric pyramid rule. It states that, for many practical networks, the number of neurons follows a pyramid shape, with the number decreasing from the input toward the output. The network construction starts using few numbers of neurons. Once the appropriate criteria have been chosen to assess network performance, this is trained and tested and its performance is recorded. Then, the process is repeated iteratively by slightly increasing the number of hidden neurons. Another critical parameter to determine is the number of training cycles. Since there is no general rule, it is necessary to implement an iterative process to identify the best configuration, even in this case.

2.4 Random Forest (RF)

The last parametric approach considered in this work is the random forest. In this case, there are two fundamental parameters to be defined: the number of trees and the maximal depth. The more the trees in the forest, the more robust (high accuracy) the cost model. However, an increased number of trees can lead to a higher computational burden. The generalisation error converges as the number of trees increases, meaning that the estimation accuracy cannot be increased after reaching a certain point. To determine the optimal value for such parameters, default values were initially defined. Once the variation ranges were chosen, it was possible to implement an iterative process to determine the optimal configuration in that given range. The default value for the number of trees was 500, while the maximum depth was 20 (Rockwell et al., 1975). Once the optimal parameters are known, the final model can be obtained. Its accuracy is evaluated using MAPE and relative percentage error.

2.5 Validation

Once created the previously presented cost models, the test phase began. As mentioned above, the initial database is divided into two broad groups: training data and testing data. The former is used to generate cost models, while the latter is used for their validation. This stage is essential in the cost estimation because it allows the understanding of which models are most suitable for the study type. Relative errors are used as metrics. There are different relative error types; the most used is MAPE (Mean Absolute Percentage Error).

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{\left|y_i - \hat{y}i\right|}{y_i}\right)\times 100\right) \tag{1}$$

Where:

- y: Known actual cost or should cost value;
- $\hat{y}$: Estimated actual cost or should cost value;
- n: Total number of pairs.

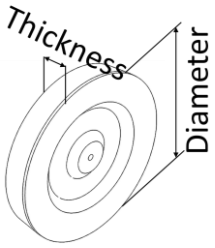
3 CASE STUDY

This work aims to assess the potential of four cost estimation methods: linear regression, artificial neural network, deep learning and random forest. The components considered for this case study belongs to the disc product family of an axial compressor. Parametric methods were developed in RapidMiner® (by RapidMiner), a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. For each technique used, a process consisting of five sub-processes was implemented. The sub-processes implemented in RapidMiner are pre-processing data, training model, testing model, creating prediction and output. Initially, should cost analyses were conducted using the LeanCOST® software. The results of this analysis permitted the definition of the starting database (90 records). The records were obtained from 9 axial compressor discs, different in dimensions. Each disc can be manufactured in 5 different material types and different batch sizes. Then, by considering other

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press

materials and batch sizes, it was possible to reach 90 records. However, the defined database is not yet suitable for applying the cost estimation methods, but it requires a pre-processing phase. The number of independent variables (cost drivers) is reduced through a DB pre-processing phase to use this low number of records in ML techniques. As stated in some research in different fields (building: Elmousalami, 2019; software development: Heiat, 2002; stamping dies: Özcan et al., 2014), the number of independent variables must be congruent with the number of records available in the database (a small DB implies few numbers of cost drivers). This initial operation was done to identify each independent variable's importance and eliminate any abnormal values or duplicate data. In this respect, a significance check analysis was carried out to determine the independent variables most closely related to the dependent one. The significance check also permitted to eliminate the independent variables that are significantly correlated with each other. The input variables (rows in the first column) and output variables (second and third columns) are summarised in Table 1, which also shows the final cost drivers chosen to implement the methods (X values) and an image of the disc (fourth column). They were distinguished between those necessary to obtain the process and raw material cost.

Table 1 cost drivers

Input	Material Cost (Cmat)	Process Cost (Cpro)	Disc Image
Final diameter (d)	X	-	
Final thickness (s)	X	-	
Material type (M1, M2, M3, M4 or M5)	X	X	
Number of slots (Ns)	-	-	
Roughness (R)	-	-	
Presence of treatments (Tratt)	-	-	
Batch size (B)	-	X	
Starting raw type (closed-die forging (Fc) or open-die forging (Fo))	X	X	
Equivalent disc weight (Peq),	X	X	
Semi-finished weight (Psl),	-	-	
Raw weight (Pg)	-	-	
Equivalent disc volume (Veq)	-	X	

Once completed the significance check analysis, anomalous and duplicate records were identified and eliminated. For process cost, the final database has the same number as the initial database. For the material cost, the number of rows has been considerably reduced (from 90 to 33). Subsequently, to avoid problems in applying the methods, material and starting raw variables were converted into dummy variables. Dummy variables could assume only two values: 0 or 1. Taking as example a disc in material M1 and manufactured by closed die forging (starting raw type: Fc), then for this component, the variables M1 and Fc will be equal to 1, while all the other variables related to material type (M2, M3, M4 and M5) and connected to starting raw type (Fo) will be null (0). Finally, the data were split into training and testing sets. In material cost, 80% of data was used as training and the remaining 20% as testing. For process cost, the percentages were 70% for training and 30% for testing. More training data were considered for material cost estimation because the database is smaller than process cost, and the materials examined have very different unit costs. Once the data subsets from training and testing have been defined, the various parametric approaches have been implemented.

The first method analysed is linear regression. The resulting equations (equations 2 and 3) are:

$$Cmat=25329,6-4621,9*M1-3640,7*M2+7858,0*M3+4104,6*M4-3697*M5+167,6*Fo-167,6*Fc-d*35,9-s*157,6+Peq*145,9 \quad (2)$$

$$Cpro=1195,0-689,8*M1-475,7*M2+906,1*M3+696,4*M4-436,9*M5+121,1*Fo-121,1*Fc-B*15,5-35476*Veq-1,3*Peq \quad (3)$$

The equations' accuracy is given by the coefficient R2, which was 0.90 for material cost and 0.84 for process cost. As the values are acceptable, the models obtained are considered valid.

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press

The second approach used was the neural network. The two main parameters to define were the number of hidden neurons and training cycles. Eight hidden neurons were considered. Through an optimisation function and respecting the relative error, it was possible to define the optimal value of training cycles. The optimum value in the training phase was 670 cycles for the material cost, with an error of 16.4%. In comparison, for process cost, it was 990 cycles with an error of 13.1%.
The third technique analysed was the deep learning. In this case, the number of hidden neurons cannot be determined by a rule of thumb. Still, it requires the use of a parametric optimisation function. Although deep learning can have more than two hidden layers, in this study, we considered only two layers with several neurons from 5 to 50 (per layer). The results of this optimisation in the training phase were:

- material cost model: 5 neurons in layer 1 and 30 neurons in layer 2, with a relative error of 24%;
- process cost model: 40 neurons in layer 1 and 5 neurons in layer 2, with a relative error of 10%.

In this case, it is unnecessary to use parametric optimisation to obtain the optimal number of training cycles. They are determined independently by stochastic gradient descent.
The last model implemented was the random forest. In this case, the parameters obtained from parametric optimisation are the number of trees and the maximal depth. Regarding the material cost, the lowest error in the training step was obtained with 50 trees and a depth of 10 (the relative error was about 23.5%). The most accurate estimate in the case of process cost is obtained with 100 trees and a depth of 10 (the relative error was about 12.0%).

4 RESULTS AND DISCUSSION

To validate and compare the parametric cost models, two testing configurations have been considered. The first one consisted of taking the testing records of the original database. The data of the first test derives from splitting the whole database in training and test records. The second configuration has been performed considering components with dimensions beyond those of the training range. In particular, a ticker component has been taken. The different parametric cost modelling techniques have been compared by considering the MAPE. Figure 2 shows the results for the cost models developed for estimating both material and process costs. The graph used for this comparison is a box-and-whisker plot. Rectangles are called "boxes", while the lines that extend vertically are called "whiskers". The boxes represent the interquartile range, which is the difference between the third quartile (upper border) and the first quartile (lower border). Within the interquartile range, 50% of the observations fall (therefore, the most frequent values are contained in this range). The line inside the boxes corresponds to the median or the second quartile. Instead, whiskers correspond to the minimum value (lower whisker) and the maximum value (upper whisker) observed after excluding the outliers.

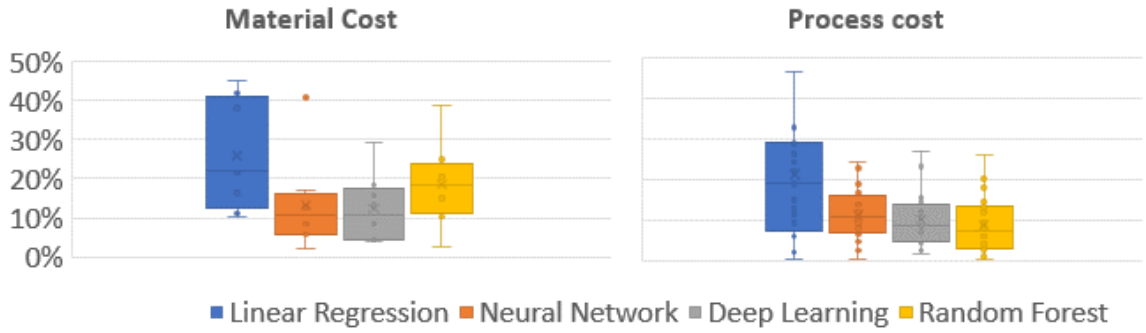


Figure 2: MAPE for the different parametric cost modelling techniques (test 1).

Considering the average performance for both material and process costs, the machine learning techniques perform better (MAPE = 12.3%, 11.2% and 13.8% respectively for ANN, DL and RM) than the traditional linear regression one (MAPE = 23%). DL appears to be the best method (MAPE = 12.2%) for estimating the material cost, whereas RF is the best for process cost (MAPE = 8.5%). The accuracy in estimating material cost is lower than the process cost. The reasons for this difference are attributable to:

- Different number of records of the two initial databases (33 vs 90, respectively for material and process cost);
- Nature of the data: very different materials (in terms of unit cost) were considered.

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press

The test data for the process costs are more numerous than the material one. This difference was the consequences of the whole data available and different independent variables considered for each cost model. To be noted that the exclusion of parameters from a database, because of their low sensitivity on the cost (i.e., the batch size for material cost), determines the generation of duplicate records that must be removed. This difference between the two databases implies more significant results reliability for the "process cost" than the "material cost". A second comparison was made to evaluate the parametric techniques' reliability in estimating the cost of components that dimensions and characteristics are beyond the range of data used for the training. For the material cost, since the low quantity of training records, reliability cannot be considered adequate. For process cost (Figure 3), instead, it is possible to observe that ML techniques behave worse than parametric ones. This conclusion leads authors to extend the type of components considered and the size of the training database.

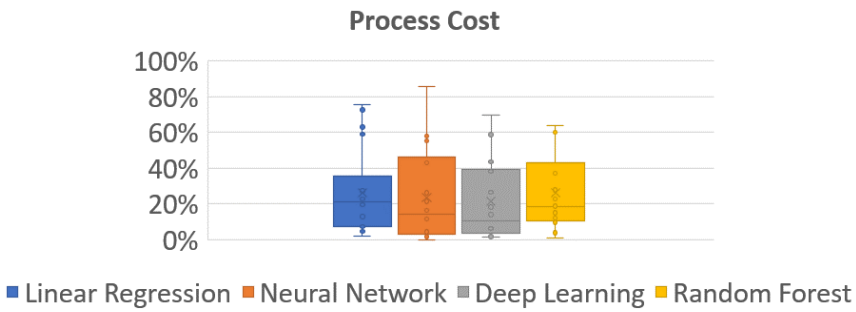


Figure 3:. MAPE for the different parametric cost modelling techniques (test 2).

The third test consisted of extending the database by should-costing additional configurations of discs, considering five batch sizes and three different materials. The new database contains around 500 records. The results obtained using this database (test 3, Figure 4) are much better than those of test 1 (Figure 2). All four parametric cost estimation methods are more accurate than test 1 (overall, MAPE was reduced by around 50%). The average MAPE reduction for material and cost estimation was -39%, -41%, -48% and -67%, respectively, for CER, ANN, DL and RF. RF is the best parametric cost modelling technique for estimating both the material and process costs, with an accuracy (MAPE) of 5.6% and 3.6%, respectively.

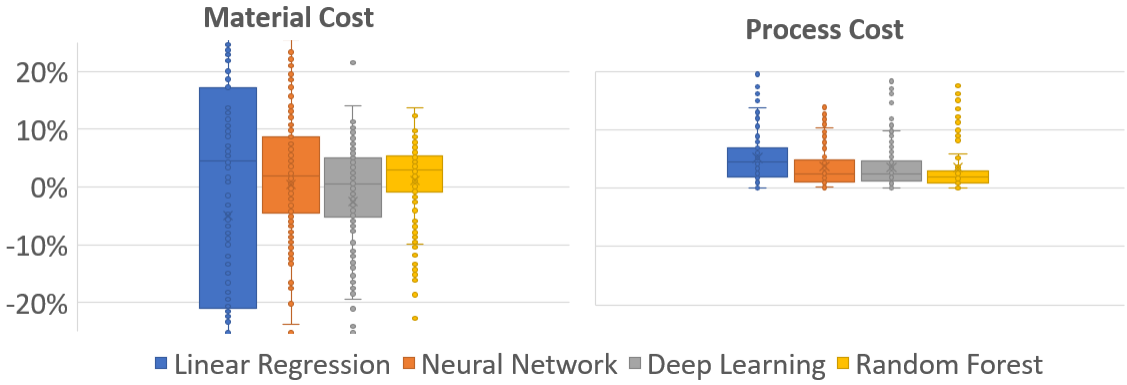


Figure 4:. MAPE for the different parametric cost modelling techniques (test 3).

5 CONCLUSIONS

The work presented in this paper originated from two research questions, "*Is it possible to overcome the historical data limitations?"* and "*What's the best parametric cost estimation approach for industrial components*?". In this regard, the goal consisted of evaluating and comparing different parametric cost estimation techniques by using non-historical data obtained from a software tool for analytical cost estimation. An original database of ninety records related to a specific family of components (i.e., axial compressor discs) was used for this goal. From the analysis carried out in this work, it is possible to conclude that the ML techniques are a valid alternative to the traditional linear

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press

regression ones. ML techniques, given their non-linearity, can manage very complex problems. For regression, it is possible to obtain greater precision if the problem is simplified (for example, by considering only one material instead of five). Regarding the differences between random forest and neural networks, the former is faster (more immediate parametric optimisation process). However, in estimating data outside the training range, Neural Networks have proved to be more accurate in most cases than the Random Forest. The accuracy in cost estimation improves while increasing the database dimensions. For a database of around 500 records of different components of the same family, Randon Forest accuracy (MAPE) is 3.6% and 5.6%, respectively, for process and raw material cost, lower than Deep Learning (3.7% and 8.0%). The cost models developed for the case study used in this paper refer to a specific family of components (typical of configurable products). Hence, they are applicable since the overall shape and manufacturing process will remain the same. This application is the main limitation of this work. Therefore, the evaluation and comparison of these cost models for non-configurable components should be investigated in the future. Furthermore, linear and non-linear methods should be implemented for other turbomachinery components (i.e., nozzles) for evaluating their performances. Besides, the Monte Carlo method could be of support to generate a more robust database of actual cost data.

REFERENCES

Abdelali, Z., Mustapha, H., Abdelwahed, N. (2019) "Investigating the use of random forest in software effort estimation", *Procedia Computer Science*, Vol. 148, pp. 343–352. https://doi.org/10.1016/j.procs.2019.01.042

Arundacahawat, A., Roy, R., Al-Ashaab, A. (2013), "An analogy based estimation framework for design rework efforts" *Journal of Intelligent Manufacturing*, Vol 24, pp. 625–639. https://doi.org/10.1007/s10845-011-0605-6

Bertoni, A., Bertoni, M. (2018), "PSS cost engineering: A model-based approach for concept design", CIRP Journal of Manufacturing Science and Technology, Vol. 29, Part B, pp. 176-190. https://doi.org/10.1016/j.cirpj.2018.08.001

Bertoni, A., Hallstedt, S. I., Dasari, S. K., Andersson, P. (2020), "Integration of value and sustainability assessment in design space exploration by machine learning: an aerospace application", *Design Science*, Vol. 6. https://doi.org/10.1017/dsj.2019.29

Bilal, M., Oyedele, L.O. (2020), "Guidelines for applied machine learning in construction industry—A case of profit margins estimation", Advanced Engineering Informatics, Vol. 43, p. 101013. https://doi.org/10.1016/j.aei.2019.101013

Boothroyd, G., Dewhurst, P., Knight, W.A. (2011), *Product Design For Manufacture and Assembly 3rd Edition*, CRC Press.

Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45, pp. 5–32. https://doi.org/10.1023/A:1010933404324

Cavalieri, S., Maccarrone, P., Pinto, R. (2004), "Parametric vs. neural network models for the estimation of production costs: A case study in the automotive industry", *International Journal of Production Economics*, Vol. 91 No. 2, pp. 165-177. https://doi.org/10.1016/j.ijpe.2003.08.005

Chen, X., Huang, J., Yi, M. (2020), "Cost estimation for general aviation aircrafts using regression models and variable importance in projection analysis", *Journal of Cleaner Production*, Vol. 256, pp. 120648. https://doi.org/10.1016/j.jclepro.2020.120648

Chou, J.S., Tai, Y., Chang, L.J. (2010), "Predicting the development cost of TFT-LCD manufacturing equipment with artificial intelligence models". International Journal of Production Economics, Vol. 128 No.1, pp. 339–350. https://doi.org/10.1016/j.ijpe.2010.07.031

Cook, R.D., Weisberg, S., (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall

Dogan, A., Birant, D., (2021), "Machine learning and data mining in manufacturing", *Expert Systems With Applications*, Vol. 166, p. 114060. https://doi.org/10.1016/j.eswa.2020.114060

Duran, O., Maciel, J., Rodriguez, N. (2012), "Comparisons between two types of neural networks for manufacturing cost estimation of piping elements", Expert Systems with Applications, Vol. 39 No. 9, pp. 7788-7795. https://doi.org/10.1016/j.eswa.2012.01.095

Elmousalami, H.H. (2019), "Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction", Computer Science, Mathematics, Engineering, Vol. abs/1909.11637. https://doi.org/10.1109/tem.2020.2972078

Favi, C., Germani, M., Mandolini, M., (2016), "Design for Manufacturing and Assembly vs. Design to Cost: Toward a Multi-objective Approach for Decision-making Strategies During Conceptual Design of Complex Products", *Procedia CIRP*, Vol. 50, pp. 275-280. https://doi.org/10.1016/j.procir.2016.04.190.

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press

Green, S.B., (1991), "How many subjects does it take to do a regression analysis?" *Multivariate behavioral research*. Vol. 26 No. 3, pp. 499-510. https://doi.org/10.1207/s15327906mbr2603_7
Heiat, A. (2002), "Comparison of artificial neural network and regression models for estimating software development effort", Information and Software Technology, Vol. 44 No. 15, pp. 911-922. https://doi.org/10.1016/S0950-5849(02)00128-3
Hihn, J., Menzies, T. (2015), " Data Mining Methods and Cost Estimation Models. Why is it so hard to infuse new ideas?", 30th IEEE/ACM International Conference on Automated Software Engineering Workshop, pp. 5-9, https://doi.org/10.1109/ASEW.2015.27
Isıklı, E., Aydın, N., Bilgili, L., Toprak, A. (2020), "Estimating fuel consumption in maritime transport", *Journal of Cleaner Production*, Vol. 275, p. 124142. https://doi.org/10.1016/j.jclepro.2020.124142
Langmaak, S., Wiseall, S., Bru, C., Adkins, R., Scanlan, J., Sobester, A. (2012), "An activity-based-parametric hybrid cost model to estimate the unit cost of a novel gas turbine component", *International Journal of Production Economics*, Vol. 142 No. 1, pp. 74-88. https://doi.org/10.1016/j.ijpe.2012.09.020
Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J.C. W., van den Bossche, P., Van Mierlo, J., Omar, N. (2018), "*Random forest regression for online capacity estimation of lithium-ion batteries*", Applied Energy, Vol. 32, pp. 197-210. https://doi.org/10.1016/j.apenergy.2018.09.182
Loyer, J.A., Henriques, H., Fontul, M., Wiseall, S. (2016), "Comparison of Machine Learning methods applied to the estimation of manufacturing cost of jet engine components". *International Journal of Production Economics*, Vol. 178, pp. 109-119. https://doi.org/10.1016/j.ijpe.2016.05.006
Mandolini, M., Campi, F., Favi, C., Germani, M., Raffaeli, R. (2020), "A framework for analytical cost estimation of mechanical components based on manufacturing knowledge representation" *International Journal of Advanced Manufacturing Technology*, Vol. 107(3-4), pp. 1131–1151.
Martinelli, I., Campi, F., Checcacci, E., Lo Presti, G.M., Pescatori, F., Pumo, A., Germani, M., (2019), "Cost Estimation Method for Gas Turbine in Conceptual Design Phase", *Procedia CIRP*, Vol. 84, pp. 650-655. https://doi.org/10.1016/j.procir.2019.04.311
Masel, D.T., Dowler, J.D., Judd, R.D (2010), "Adapting Bottoms-up Cost Estimating Relationships to New Systems", *ISPA/SCEA Joint Annual Conference and Training Workshop*.
Niazi, A, Dai, J.S., Balabani, S., Seneviratne, L. (2005), "Product cost estimation: technique classification and methodology review", *Journal of Manufacturing Science and Engineering*, Vol. 128 No. 2, pp. 563–575. https://doi.org/10.1115/1.2137750
Ning, F., Shi, Y., Cai, M., Xu, W., Zhang, X. (2020), "Manufacturing cost estimation based on the machining process and deep learning method", *Journal of Manufacturing Systems*, Vol. 56, pp. 11-22. https://doi.org/10.1016/j.jmsy.2020.04.011
Özcan, B., Fığlalı, A. (2014), "Artificial neural networks for the cost estimation of stamping dies", *Neural Computing and Applications*, Vol. 25, pp.717-726. https://doi.org/10.1007/s00521-014-1546-8
Rockwell, R.C., (1975), "Assessment of multicollinearity: The Haitovsky test of the determinant." *Sociological Methods & Research*, Vol. 3 No. 3, pp. 308-320. https://doi.org/10.1177/004912417500300304.
Sala, R., Zambetti, M., Pirola, F., Pinto, R. (2018), "How to select a suitable machine learning algorithm: A feature-based, scope-oriented selection framework", *23rd Summer School "Francesco Turco"-Industrial Systems Engineering 2018*, Vol. 2018, pp. 87-93.
Tan, H., Wang, H., Chen, L., Shi, F. (2011), "Dummy Variable Model Analysis With Law Factors on Safety Production in Chinese Coal Mine Industry", *Procedia Engineering*, Vol. 26, pp. 2383-2390. https://doi.org/10.1016/j.proeng.2011.11.2449
Ulrich, K., Eppinger, S.A. (2011), *Product Design and Development*, McGraw-Hill Education, New York.
Verlinden, B., Duflou, J.R., Collin, P., Cattrysse, D. (2007), "Cost estimation for sheet metal parts using multiple regression and artificial neural networks: A case study", *International Journal of Production Economics*, Vol. 111 No. 2, pp. 484-492. https://doi.org/10.1016/j.ijpe.2007.02.004
Wang, H.S., Wang, Y.N., Wang, Y.C. (2013), "Cost estimation of plastic injection molding parts through integration of PSO and BP neural network", *Expert Systems with Applications*, Vol. 40 No. 2, pp. 418-428. https://doi.org/10.1016/j.eswa.2012.01.166
Weichert, D., Link, P., Stoll, A., Ruping, S., Ihlenfeldt, S., Wrobel, S. (2019), "A review of machine learning for the optimization of production processes", *The International Journal of Advanced Manufacturing Technology*, Vol. 104, pp. 1889–1902. https://doi.org/10.1007/s00170-019-03988-5

https://doi.org/10.1017/pds.2021.499 Published online by Cambridge University Press