Damir Kolich
Nikša Fafandjel
Y. Lawrence Yao

ISSN 0007-215X
eISSN 1845-5859

DATA MINING METHODOLOGY FOR DETERMINING THE OPTIMAL MODEL OF COST PREDICTION IN SHIP INTERIM PRODUCT ASSEMBLY

UDC 629.5(05):519.6:519.25:691.81
Original scientific paper

Summary

In order to accurately predict costs of the thousands of interim products that are assembled in shipyards, it is necessary to use skilled engineers to develop detailed Gantt charts for each interim product separately which takes many hours. It is helpful to develop a prediction tool to estimate the cost of interim products accurately and quickly without the need for skilled engineers. This will drive down shipyard costs and improve competitiveness. Data mining is used extensively for developing prediction models in other industries. Since ships consist of thousands of interim products, it is logical to develop a data mining methodology for a shipyard or any other manufacturing industry where interim products are produced. The methodology involves analysis of existing interim products and data collection. Pre-processing and principal component analysis is done to make the data "user-friendly" for later prediction processing and the development of both accurate and robust models. The support vector machine is demonstrated as the better model when there are a lower number of tuples. However as the number of tuples is increased to over 10000, then the artificial neural network model is recommended.

Key words: *data mining; pre-processing; principal component analysis; support vector machine regression; artificial neural network; shipbuilding*

1. Introduction

Data mining is used in various industries from pharmaceutical to real estate and also the manufacturing industry to varying levels and degrees. In manufacturing it is represented in CNC machining as well as in steel galvanization industries. It is therefore an interdisciplinary field, which makes the use of appropriate algorithms and techniques in order to predict, classify or determine relationships between information that would otherwise be unknown according to Agard and Kusiak [1]. However, the use of data mining to accurately predict the costs of interim product assembly is lacking. Panel and micro-panel assembly lines are typical production processes in most new-building shipyards, since panels make up to 70 per cent of

the steel weight of ships. These interim products vary in size, weight and dimensions from one another. Therefore, many shipyards assemble panels on up to four different production lines or stations. These include the large panels, robotically assembled micro-panels, automatically assembled micro-panels and manually assembled micro-panels. These interim products are also more frequently sub-contracted. Therefore, having a method to quickly and accurately predict the costs of interim products, means that it will be possible to check whether sub-contractors are realistic with their bids. The approach and practice is valuable since it will drive down shipyard costs by reducing the engineering man-hours. It takes about four hours to produce an accurate Gantt chart of a typical interim product. Since a typical ship contains over two thousand interim products, this translates to about 10000 man-hours. If the rate of an experienced engineer is 50 Euros, this means 500,000 Euros in savings per ship.

The purpose of the data mining methodology described in this paper is to demonstrate and explain how existing data about interim products can be used to predict the costs of future interim products. The key steps include data analysis, determining what pre-processing is necessary, and the testing of pertinent regression models to determine which data mining technique would produce the best model for making cost predictions. Up to now most cost predictions are done without the necessary pre-processing and sometimes result in an unreliable model, where there are a multitude of predictors with different units and magnitudes. In this work however, pre-processing to ensure robustness in addition to testing different regression models to find the best model is a new approach.

Pre-processing makes the inputted data user friendly for further analysis. The pre-processing includes the standard centering, scaling, as well as correlation analysis and principal component analysis (PCA). Correlation analysis provides the correlations between the predictors as well as between each predictor and the outcome. The authors tried out basic linear regression, which yielded coefficient of determination (R^2) and root mean square errors (RMSE), which do not appear to be optimal (Kuhn and Johnson 2013) [15]. Then a principal component regression (PCR) model, which used the pre-processed data, was applied as well. The hypothesis of the authors is that since most industrial processes exhibit nonlinear behavior according to Kim and Lee [11], the assembly costs could be best estimated by a non-linear predictive method suggested by Ren et al. [23] such as the support vector machine (SVM) or the artificial neural network (ANN). The SVM has advantages in that it finds a global minimum, while the ANN finds multiple local minima according to Lamorski et al. [16]. The authors Kuhn and Johnson [15] created a table of all the different regression techniques with R^2 and RMSE values.

The practical application is making use of the shipyard database of interim products and the characteristics or predictors of each interim product, which includes panel mass, steel plate thickness, panel length, panel width, stiffener height, stiffener thickness, stiffener length, stiffener number, stiffener number of types, stiffener tensile strength, and steel plate tensile strength. This is a total of eleven characteristics or predictors, whereas the outcome that is of interest is cost. Therefore in a shipyard database the characteristics listed above are readily available in addition to the outcome cost for interim products already assembled. However, since each new ship has different interim products and no two interim products are exactly the same from ship to ship, it would be helpful to make use of the eleven characteristics of each interim product along with the known cost to develop an accurate mathematical model that will predict costs for future new interim products for ships that are yet to be built. This way the shipyard management could precisely know what the costs are of building the future interim products and make a more accurate estimate for the cost of the entire ship. Data mining prediction methods are crucial due to the plethora of interim products built in a shipyard every year which numbers in the tens of thousands, and therefore any industry which assembles interim products such as modular construction of buildings or bridges and even

airplanes could be very accurate and efficient in predicting the costs of each interim product and through summing up, the entire cost of the final product.

2. Literature review

The cost estimations that are performed in shipyards and other large manufacturing companies were done exclusively on the ship or the final product, which was compliant to the traditional work breakdown structure (WBS). However, since the transformation of shipyards from fabrication manufacturing to assembly of interim products is becoming more complex, it has become necessary to break down the assembly costs of a ship into its building blocks. This is in line with the product work breakdown structure (PWBS) suggested by Kolich et al. [12,13]. By accurately determining the costs of assembling all the interim products of a ship, it is therefore more realistic to calculate the total cost of building the entire ship.

The traditional method for determining the cost of an interim product is to develop a detailed Gantt chart for each interim product from which it is possible to calculate the production times and man-hours and therefore the costs explained by Storch et al. [25]. While very accurate, the drawback to developing a detailed Gantt chart for each interim product is that it would require an experienced production planner at least 4 hours per interim product. Since a typical commercial vessel consists of thousands of interim products this translates to 10000 or more engineering man-hours.

Trumbule et al. [28] have developed the Product Oriented Design and Construction (PODAC) cost model, which utilizes multiple modules to make estimates of production cost. Likewise the cost assessment method developed by Caprace et al. [3] also requires the integration of multiple databases into one system to accurately predict the costs of new interim products. These include a scheduling database, a CAD/CAM database and a rules database. These interfaces appear to approach complexity and are susceptible to modules, which need to be simultaneously updated.

The paper by Agard and Kusiak [1] uses data mining clustering techniques and applies them to the design of product families which is very beneficial in standardizing designs. Whereas some of the techniques from this work would be helpful in determining interim product standardization for ships, it is not possible for cost prediction. Another clustering paper was developed by the authors' earlier involving data mining clustering to aid in grouping micro-panel interim products to be assembled in a combined assembly line which combines the best assets of both production lines to yield a more efficient one explained by Kolich et al. [14]. Likewise, Osma [21] developed an analytical approximation for the stress-strain relationship for steel. However, the polynomials that were developed in this paper are not relevant for cost prediction of ship interim products, which are more complex and harder to predict. For instance, the SVM or ANN need to be used which cannot be represented by polynomials.

Fafandjel et al. [6] have developed a model to estimate the production costs of the end product, a particular ship type. Other research has focused on the Work Breakdown Structure cost estimation recommended by Nan et al. [20]. The method takes the triangular distribution approach with the least, maximum and most probable creating a distribution of values for each task using hypercube sampling. This method does not permit for the cost estimations of the many different types of interim products and variations. It is good for initial and general estimation of shipbuilding task costs. However the above works do not model interim product cost prediction. Jin et al. [10] discusses the use of digital manufacturing technology for predicting manufacturing costs. Since the level of accuracy of the digital manufacturing technology was not demonstrated in the paper and much of industry has not invested in this technology, it is necessary to have alternative and practical methods for cost prediction.

Stockton et al. [26] developed a cost model for manufacturing processes using data mining. The authors treated linear regression and ANN. However, the shortcomings are that the predictors were not pre-processed. Likewise, the ANN model was not as powerful as it could have been had it been fine-tuned, as is done by the authors of this paper. Buddharaju et al. [2] developed an SVM model to predict the costs of building a new well during drilling for petroleum. They determined that the SVM model is better than the ANN model due to ANN over fitting tendencies. However, the authors with the exception of data normalizing did not analyse or discuss the pre-processing necessary in order to ensure a robust model. Likewise the shortcomings of the SVM model are that while accurate, it is not able to predict models with more than 10000 tuples of data, which could be the case. Therefore, having a second accurate method for estimating costs would be necessary for future models, where it is likely that the database will contain many tuples of data.

Das et al. [5] found that the SVM model was better for prediction of "field hydraulic conductivity of clay liners" than the ANN model by using RMSE and R^2 as criteria for comparison. The SVM was likewise better than the ANN for prediction of "energy loss in soil working machines" according to Taghavifar and Mardani [27] and for prediction of pedotransfer functions (PTFs) "which estimate soil hydraulic parameters" Lamorski et al. [16] using the same criteria as mentioned above. On the other hand, Mansfield et al. [17] found that the ANN model performed better than the SVM model in predicting wood properties. Since ANN and SVM are both powerful prediction algorithms, it is not possible to automatically conclude which method is better. Likewise it is necessary to note that the authors of the above papers did not discuss correlation analysis or PCA, both of which are important pre-processing techniques that ensure a robust model. Therefore it is necessary to undertake the steps and follow the methodology described in this paper. Likewise, it is necessary to be able to predict the cost of individual interim products since shipyards also purchase interim products from other manufacturers as well as sell interim products to other shipyards depending on the market. This flexibility of buying and selling is advantageous in order to be competitive and an accurate cost prediction model will help management to both buy and sell at realistically competitive prices.

3. Background

Data mining is the application of pertinent algorithms to data from databases so as to intelligently analyse and make it practical and useful for future uses. It is used in the mining of five different types of knowledge: 1) concept description, 2) association, 3) classification, 4) clustering and 5) predicting according to Choudhary et al. [4]. The progressive increase in the use of computers and digital technologies from three-dimensional product design to computer numerically controlled machines, more data is available than ever before in various manufacturing databases as explained by Han and Kamber [8]. For the purposes of this paper the last type of knowledge is applied since it solves the interim product assembly cost prediction. The prediction of the performance of manufacturing processes is important in order for a company to determine the cost of producing the interim product, according to Feng et al [7].

3.1 Analysis and pre-processing of data

Prior to performing prediction of costs, it is necessary to collect the data and then to analyse it in order to determine what if any pre-processing is necessary. According to Mikovsky et al. [19], data pre-processing is a prerequisite to performing quality data mining algorithms. The predictors of data can be expected to have certain levels of correlation as well as differences in magnitude and scale, and if the data is not pre-processed properly this could

have adverse affects on the algorithms used to make predictions. Yan et al. [29] stress the importance of pre-processing data and ranking quality above the quantity of data used. Likewise according to Razavi et al. [22], pre-processing takes up to 80% of the total time in the knowledge discovery in databases (KDD), which is a common synonym for data mining.

In order to resolve the fact that some predictors have variance many times larger than others of the same tuple, it is necessary to centre and scale all of the predictors. Otherwise those predictors with higher orders of magnitude will dominate and the trained model will not be accurate, and therefore this was done in this paper.

3.1.1 Covariance and correlation checking of data

This is the first work where the importance of robustness is demonstrated and applied in a shipbuilding project prediction. Robustness is directly related to eliminating predictors with high collinearity. The main reason to avoid predictors with high collinearity is that the model will be sensitive to changes as well as mistakes that could occur in the data collection process. Correlation equals covariance divided by the product of the standard deviations of the predictors. Therefore, the values are always between 0 and 1. When the correlation between two predictors is close to or equal to 1 then they are considered to be collinear. High correlations between predictors are harmful to the performance of the model. Equation 1 below demonstrates how high correlations between predictors' results in a non-robust model. There is one outcome Y, and predictors X and predictor Z are identical (correlation = 1) and a, b, and c represent coefficients.

$$Y = a + bX + cZ + error = a + bX + cX + error = a + (b + c)X + error \quad (1)$$

Neither b nor c can be estimated, only their sum (e.g. b=0.4 and c=0.6, then b+c=1). However possible estimates are b=1001 and c=-1000, which is far from the truth. When correlation is ≈ 1, then we can estimate b and c, but these estimates are very unstable (which can be seen from huge estimated standard errors for these parameter estimates), so our model becomes extremely sensitive. This work is the only one up to now which treats the importance of creating not only an accurate but a robust model in cost prediction as explained above.

3.1.2 Principal component analysis

Principal component analysis (PCA) is the cure to the problem of multicollinearity between predictors. It does this by transforming the original predictors through rotation until they become orthogonal to one another. As a result the new components have zero correlation between each other explained by Mere et al. [18]. Each component is an eigenvector that is represented by eigenvalues for all of the original predictors. In addition it is possible to decrease the number of principal components through cross validation. In this way PCA performs dimensionality and noise reduction as well as neutralizing the effects of correlated predictors according to Yuan et al [30], which is desirable in creating a model that is accurate, relatively simple and robust.

The restricted optimization problem for the first component (PC_1) is represented by equations 2 and 3 where w_1 is the vector of weights with which each original predictor enters the first component; n is the number of tuples, X is the data matrix of all predictor values and all tuples and p is the number of predictors in each tuple. T is the transpose so that the matrix matches up.

$$\overrightarrow{w_1} = \arg w_{\bar{w}} \sum_{i=1}^{n} (\vec{X}_i^T)^2, \quad (2)$$

$$\text{under the condition that} \quad \sum_{j=1}^{p} w_j^2 = 1, \quad (3)$$

Then $PC_1 = X^T \vec{w}_1$. (4)

Once the pre-processing of the predictors is completed, it is necessary to consider which model will best predict assembly costs of the interim products. As the data consists of multiple predictors, it is reasonable to assume that a linear regression method will not produce the best model since most real world models are nonlinear.

3.2 Regression methods

Prediction makes use of various regression methods. This includes simple linear regression, PCR as well as nonlinear techniques such as SVM and ANN [7, 24].

3.2.1 Principal component regression

Principal component regression is a two-step procedure. Firstly, a principal components analysis of the predictors is performed. Note that it is advised to standardise (centre and scale) each predictor of the new interim product prior to the principal components analysis, because the predictors are measured on highly different scales. Secondly, a linear regression is performed, on the newly derived principal components, not on the original predictors. Through cross validation of the training set, the optimal number of principal components to use can be determined.

3.2.2. Support vector machine regression model

The SVM regression model is the intercept β_0 plus the product of the weights α and the kernel functions $K(x_i, u)$ for all the tuples n (see Equation 5), where u are the predictor values of an interim product whose cost shall be predicted.

$$f(u) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x_i, u) \tag{5}$$

The advantage of support vectors is that through the use of only a few kernels it is possible to create complex functions. The typical kernel function used is the radial basis kernel

$$K(x_i, u) = \exp(-\sigma \bullet \sum_{j=1}^{p} (x_{ij} - u_j)^2), \tag{6}$$

where p is the number of predictors/components when PCA is performed x_{ij} is the j-th predictor value of the i-th interim product and u is as mentioned above. Sigma (σ) is a tuning parameter, which represents the width of the kernel. As sigma increases, the regression function becomes smoother.

In order to estimate the parameters, SVM regression minimizes the so-called epsilon (**ε)** insensitive loss, which is roughly similar to ordinary linear regression's squared error loss. The **ε**-insensitive loss function (L_ε) assigns no loss to residuals, which are smaller in absolute value than **ε**, and linearly increasing loss for residuals larger than **ε** in absolute value (Kuhn and Johnson 2013) [15]. **ε** is a tuning parameter related to the degree of insensitivity of the loss function towards small errors/residuals. As **ε** increases**,** the regression function becomes smoother. To control the smoothness of the regression function, a penalization term is added to the goal function. With the penalization term, the goal then is to minimize Equation 7.

$$\min_{\beta_0,;\alpha_1 \ldots . \alpha_n} \left\{ C \cdot \sum_{i=1}^{n} L_\varepsilon (y_i - \hat{y}_1) + \sum_{j=1}^{p} \sum_{i=1}^{n} \alpha_i \alpha_j x_i^2 \right\} \tag{7}$$

where C is a tuning parameter where for large C, the goal of minimizing the residuals dominates, and we get over fitting. For small C, emphasis is on smoothness of the regression

function, and then there is under fitting. The three tuning parameters σ, ε and C control different aspects of the smoothness of the regression function. They can be chosen with cross validation. The SVM regression model is trained for a large set of combinations for σ, ε and C and applied to the cross-validation held-out sets. Those values for σ, ε and C are chosen from the cross validated training set, which give best predictive accuracy in terms of the smallest RMSE value.

3.2.2. Artificial neural network

Many sources demonstrate that the SVM has higher predictive accuracy than the ANN [5]. However, there is also literature showing that the ANN performs better according to Mansfield et al. [17]. The structure of an ANN resembles that of a brain. A network of neurons is used to learn the unknown regression function. This artificial neural network has a hidden layer of neurons/nodes between the input and the output layers. Since ANNs use gradients to optimize model parameters it is necessary to perform PCA when there are high correlations between predictors. Loops are used to optimize the three tuning parameters weight decay (lambda), number of hidden units and the number of principal components to use in the model, which is a novel approach due to its efficiency and accuracy.

Equations 8 and 9 below explain the relationship between the input layer and the hidden layer, where h represents the sum of the linear equations for each hidden unit Equation 8, which is then transformed by Equation 9. There are total of k-hidden units, which are determined by cross validation. The linear equations are transformed by a sigmoidal function $g(u)$ in order to learn a complex system that cannot easily be described otherwise. Then Equation 10 is a linear equation that is the sum of the product of the weights of the hidden nodes and the corresponding actual weights. The final outcome f(x) is equal to the sum of the intercept γ_0 and the sums of the weights for each hidden unit multiplied by the output of the respective hidden unit.

$$h_k(x) = g(\beta_{0k} + \sum_{j=1}^{p} x_j \beta_{jk}), where \tag{8}$$

$$g(u) = \frac{1}{1+e^{-u}} \tag{9}$$

$$f(x) = \gamma_0 + \sum_{k=1}^{H} \gamma_k h_k \tag{10}$$

Weight decay λ is another tuning parameter used to decrease over-fitting and regularize the model [15]. Large regression coefficients must have a strong positive effect on the model accuracy to be acceptable. To find the coefficients equation 11 is minimized, where y represents the actual cost, f represents the regression function, $f(x_i)$ is the predicted cost, H is the number of hidden units, p is the number of predictors or components when PCA is used, betas β_{jk} are the parameters with which the predictors enter the hidden layers, gammas γ_k are the parameters with which the outputs of the hidden layers are combined to yield the prediction.

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \sum_{k=1}^{H} \sum_{j=0}^{P} \beta_{jk}^2 + \lambda \sum_{k=0}^{H} \gamma_k^2 \tag{11}$$

As the regularization value increases the fitted model becomes smoother and less likely to over-fit the training set. In choosing the number of hidden units they are usually smaller in number than predictors.

4. Problem description and approach

Since even among the micro-panels the range of design characteristics differs considerably, it is necessary to determine a methodology for analysing the entire scope of possible panels and micro-panels, which are assembled in a ship production environment. The design characteristics or predictors of panel interim products includes the following: weight (kg); steel plate thickness (mm); panel length (mm); panel width (mm), stiffener height (mm); stiffener thickness (mm); stiffener length (mm); stiffener number of types; stiffener number; stiffener tensile strength (N/mm^2); plate tensile strength (N/mm^2). The use of the above mentioned numerous physical characteristics is a novel approach to harness the many different types of interim products to be used as a basis to develop an accurate mathematical model for predicting future costs related to new interim products.

A set of 229 tuples developed from 45 drawings of micropanels and panels is generated, with four classes of interim products (28 CA's – automated micro-panels, 10 CR's robotic micro-panels, 4 MP's manual micro-panels and 3 P's panels). See Figure 1.

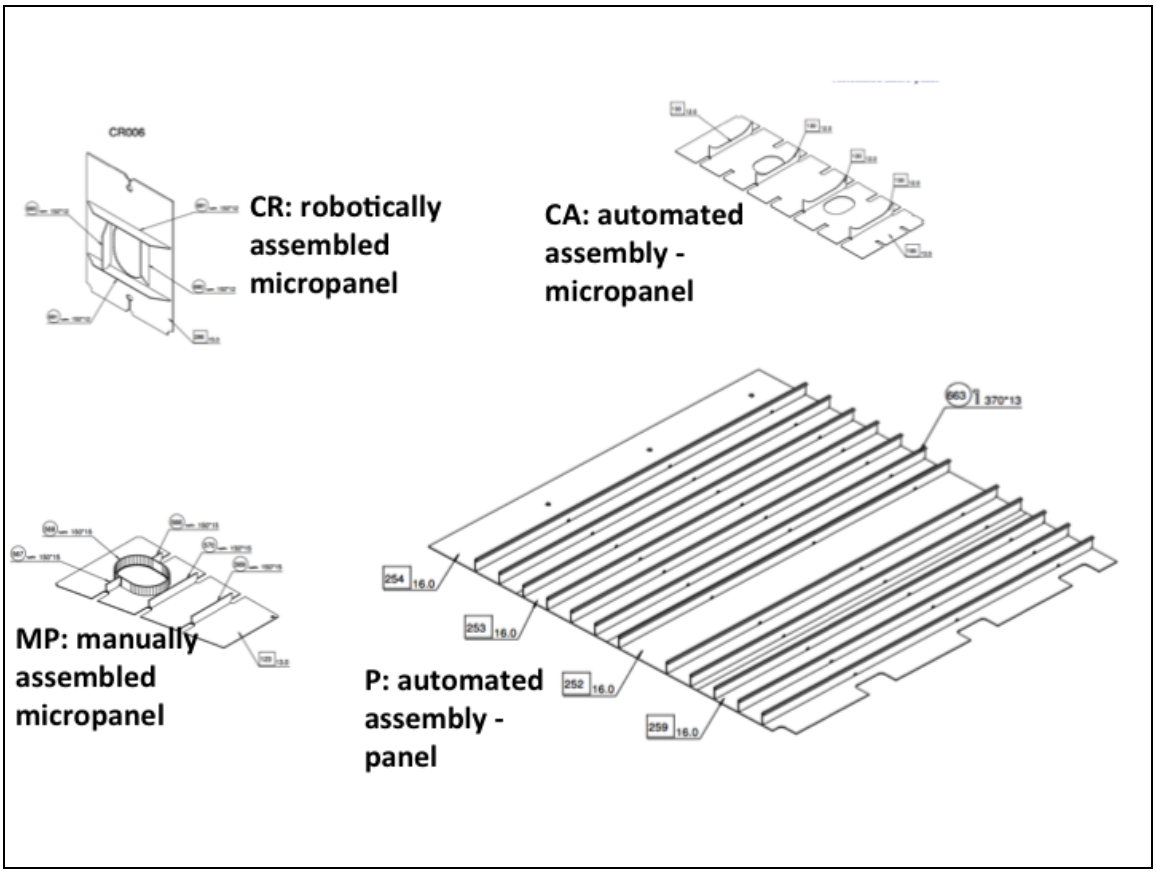

Fig. 1 Interim products assembled in the sub-assembly workshop on 4 different process lanes.

The main differences between the interim products are that the P's are the longest and widest panels, and are assembled on a panel line with greater capacity. The MP interim products are manually assembled, due to having elements, which require circular or curved welding which can only be performed by skilled welders manually. The CA's are automated assembly micropanels, which have all stiffeners positioned in one direction and allow for welding by automated welding machines located on the CA assembly line. Finally, the CR micropanels have stiffeners that need to be welded in both the longitudinal and in the transverse directions. This requires the robotic assembly line, which is designed to weld in both directions. Therefore, even though the micropanels and panels have many similarities,

the differences in the size and the configuration of stiffeners and types of elements, classify the type of interim product and therefore the corresponding assembly line. The costs of these micropanels are determined by expert analysis according to Storch et al. [25], which includes the development of Gantt charts, and interim product algorithms explained by Kolich et al. [12]. The aim of the work in this paper is to eliminate the need for expert analysis which is both time consuming and costly and develop a tailor made methodology which can be used by any shipyard which assembles interim products to predict the costs of future interim products.

The summary of steps in following the data mining methodology to choose the best model to predict interim product costs are:

(1) Analysis of the present panel interim products:
(2) Data mining analysis:

- Analysis of tuples to determine what pre-processing is necessary
- Pre-processing analysis (centring, scaling, correlation analysis and PCA)
- Linear regression, PCR
- Nonlinear regression models: ANN, SVM regression

(3) Best regression model for predicting costs of panel and micropanels assembly based on R^2 and RMSE values.

4.1 PWBS cost estimation of interim product assembly

As mentioned earlier, in order to estimate the cost of assembling a micro-panel or a panel, it is necessary to draw up a Gantt chart of the assembly activities. The Gantt chart describes all the activities, the duration time, the type and the number of workers working at that activity. For each activity, different trades are used. For example the crane operator transports the steel plates. However, ship fitters perform the tack welding, whereas welders perform welding, depending on the assembly line. Please note that the man-hour values used in this paper include the total man-hours of all trades for the assembly of each interim product. Since the sales department usually uses an average rate in its estimations it is done in this way for this paper as well. The formula for calculating the total number of man-hours, *Mhrs* is in equation 12. The duration time is *DT*, whereas the subscript *i* represents the activity number and *a*, the total number of activities and *O* is the number of operators for that activity according to Kolich et al. [13].

$$Mhrs = \sum_{i=1}^{a} (DT_i \cdot O_i) \tag{12}$$

Finally to determine the cost the *Mhrs* is multiplied by the shipyard man-hour rate, which is set by the shipyard management. This yields the cost of the interim product (See equation 13).

$$Cost = Mhrs \cdot rate \tag{13}$$

4.2 Analysis and pre-processing

Analysis of the data is needed to determine what pre-processing is necessary. Prior to even pre-processing of data, it is possible to train and test a linear regression model of the data tuples. Since the linear model is not difficult to create and test, this was performed first. Determining the best linear fit by minimizing the sum of square errors is the method of linear regression [9].

The pre-processing involves first analysing the tuples of data of interim products, which includes the number of predictors and the outcome. Therefore centring and scaling will need to be performed. Likewise a correlation analysis is performed to determine if there is a

problem of multicollinearity. This is something which is rarely performed by industry in developing prediction models, and never in the shipbuilding industry up to now.

PCA eliminates the problem of collinearity between predictors. Therefore even though a linear model and a PCR model may have similar R^2 values, the PCR model will be more robust for applications. Even when the R^2 values are similar with regards to the test set, when the model gets tested in a real life application, the PCR model will yield more accurate results due to its robustness from the elimination of multicollinearity through PCA.

Once the PCA is performed the deduction about whether the data fits in a linear or a non-linear model can be made. One model is the PCR that essentially uses the principal components that are not collinear, and therefore, the model will be stable. However, since the model is still a linear one, it is probable that the accuracy is limited. Whereas the model is robust, which means that it is not sensitive to changes in the way data is collected; the prediction results will not be optimal because a non-linear model is necessary.

4.3 Implementation of models

Therefore using the above-mentioned equations and pre-processing, it can be expected that it is necessary to both standardize and perform PCA on the predictors. Then a PCR model is created, cross validated according to Kuhn and Johnson [15] on the training set and then finally tested. Likewise the same is done using the procedures explained earlier for the SVM regression and ANN methods. It is important to use the principal components in these models to ensure robustness and reduced dimensionality, which simplifies the models while retaining accuracy.

5. Results and discussion

The predictors were analysed using scatter plots (See Fig. 2) and with correlation analysis (See Table 1). The predictors shown in Figure 2 are mass, plate thickness, panel length and panel width. A pure correlation between two predictors would be represented by a diagonal line pattern among the scatter points. Whereas there is no exact diagonal line pattern in the figure between any two predictors, there are some patterns between predictors of diagonality which demonstrate some degree of correlation. This includes the correlations between mass and panel length and mass and panel width. This makes sense because as the length and width of the panels increase so does the mass. However it is not at a correlation of one. For instance if the bottom left square is analyzed, it shows mass in kilograms on the x-axis and width in millimetres on the y-axis and the pattern has some degree of correlation. The square to the top right likewise illustrates a degree of correlation as well. The x-axis is represented by width in millimetres, whereas the y-axis is mass in kilograms. It was determined that some of the predictors have high correlations between one another, as well as different magnitudes. Therefore, pre-processing of scaling, centering and principal component transformation of the predictors was necessary.

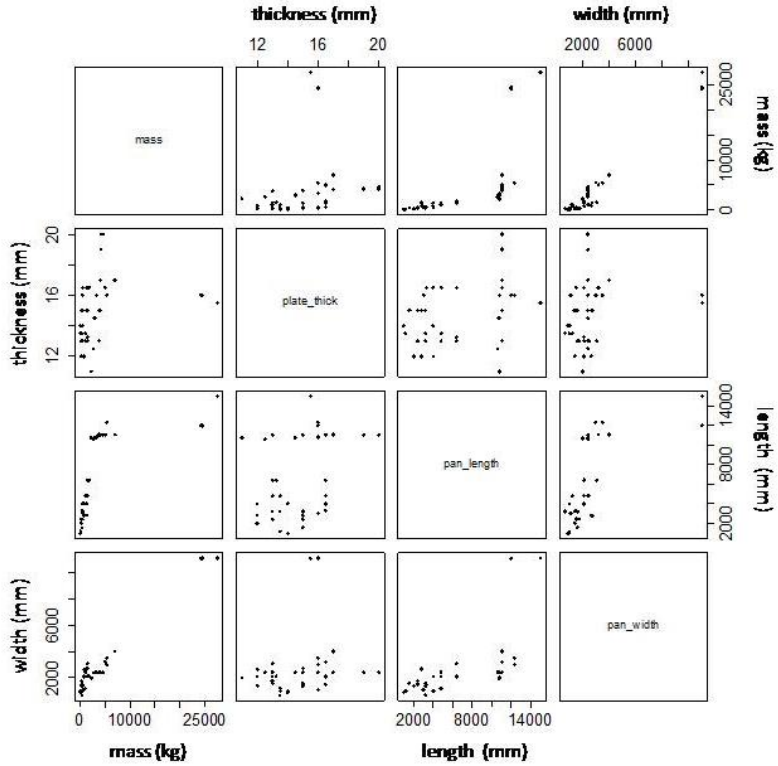

Fig. 2 Scatter plots between predictors

The basic ordinary linear regression model without any pre-processing of the data yields the following equation 14. This was done with consecutive backwards elimination of predictors by starting with the predictors with the highest p values and then working progressively down [9].

$$\cos ts = -289 - 0.2 \cdot mass + 23 \cdot plate\_thick + 0.03 \cdot pan\_length + 0.04 \cdot stiff\_length + 2.5 \cdot stiff\_no\_types - 29 \cdot stiff\_no \quad (14)$$

Finally, the R^2 value was found to be 0.795 and the RMSE is 231, which means that the model accuracy is not optimal. This was to be expected since the data is non-linear and therefore, a linear model is not able to accurately predict a non-linear situation.

5.1 Analysis of tuples

From the table below (See Table 1), it is clear that the following predictors have high collinearity: mass and panel width value of 0.989; mass and stiffener length 0.996; panel width and stiffener length value of 0.989. These are definitely approaching one. The existence of multicollinearity between predictors will lead to a detrimental linear regression model. This means that the above model in equation 19 has high collinearity between two of the predictors, mass and stiffener length and is not reliable. It is necessary to perform a PCA. This treatment of the importance of eliminating multicollinearity prior to performing data mining prediction of interim product cost is a novel approach in the development of an accurate and robust predictive model in manufacturing.

Table 1 Correlation values between the 11 predictors.

```
> cor(Predictors)
                    mass  plate_thick  pan_length    pan_width    stiff_ht  stiff_thick stiff_length
mass          1.000000000  0.379516320  0.73766616  0.989720773  0.88468161  0.007823864   0.99613915
plate_thick   0.379516320  1.000000000  0.48119209  0.346085392  0.54878274 -0.008475478   0.37856344
pan_length    0.737666158  0.481192086  1.00000000  0.713129117  0.74265333 -0.144878672   0.75306305
pan_width     0.989720773  0.346085392  0.71312912  1.000000000  0.88803469 -0.009080006   0.98985579
stiff_ht      0.884681607  0.548782739  0.74265333  0.888034689  1.00000000  0.061490643   0.88798536
stiff_thick   0.007823864 -0.008475478 -0.14487867 -0.009080006  0.06149064  1.000000000  -0.01646799
stiff_length  0.996139147  0.378563443  0.75306305  0.989855786  0.88798536 -0.016467985   1.00000000
stiff_no_types 0.099185946  0.185056415  0.20000262  0.108435531  0.16000314 -0.140143664   0.11624746
stiff_no      0.628059366  0.252093065  0.70005620  0.644183052  0.55047287 -0.101610431   0.65677340
stiff_ten_str 0.018103735 -0.076239797 -0.02562691  0.014853136 -0.01554613  0.015137039   0.02149205
plate_ten_str 0.013672641 -0.030858650 -0.02881212  0.008667471 -0.01337658 -0.016141323   0.01638297
              stiff_no_types     stiff_no stiff_ten_str plate_ten_str
mass              0.099185946  0.628059366   0.018103735   0.013672641
plate_thick       0.185056415  0.252093065  -0.076239797  -0.030858650
pan_length        0.200002622  0.700056196  -0.025626910  -0.028812116
pan_width         0.108435531  0.644183052   0.014853136   0.008667471
stiff_ht          0.160003141  0.550472868  -0.015546132  -0.013376579
stiff_thick      -0.140143664 -0.101610431   0.015137039  -0.016141323
stiff_length      0.116247462  0.656773396   0.021492054   0.016382967
stiff_no_types    1.000000000  0.157023855   0.066007130   0.009397587
stiff_no          0.157023855  1.000000000  -0.006519384  -0.002295265
stiff_ten_str     0.066007130 -0.006519384   1.000000000   0.378951372
plate_ten_str     0.009397587 -0.002295265   0.378951372   1.000000000
```

5.2 PCR model

Through cross-validation of the training set, it is determined that the PCR linear model has four principal components (See Fig. 3). Equation 15 represents the PCR model.

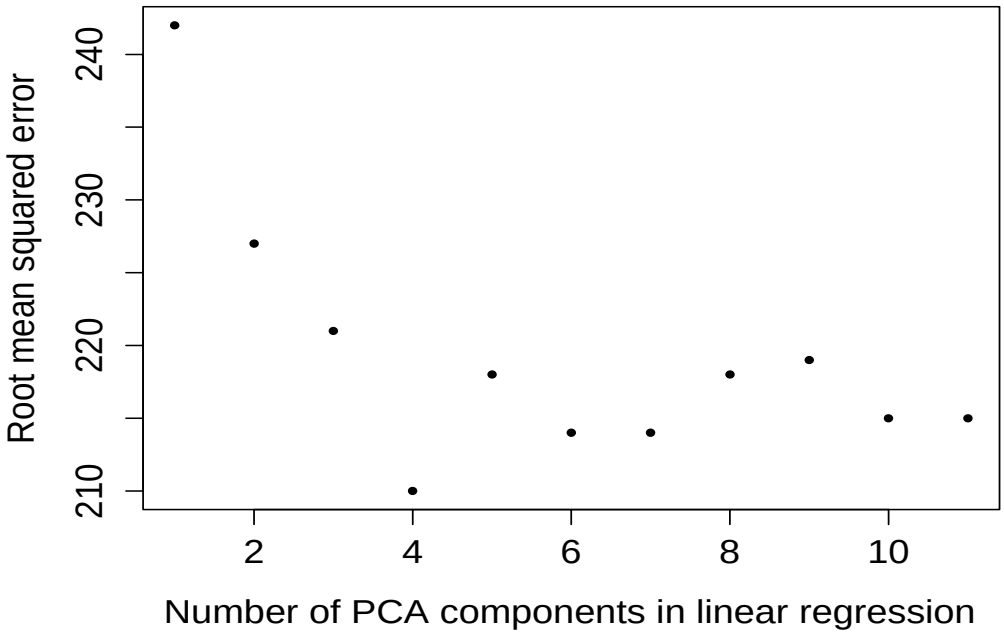

Fig. 3 Root mean square error values of Principal Components in PCR model.

$$\begin{aligned} costs = &-179 + 0.016 \cdot mass + 17.06 \cdot plate\_thick + 0.007 \cdot pan\_length \\ &- 0.001 \cdot pan\_width + 1.5 \cdot stiff\_ht - 15.7 \cdot stiff\_thick + 0.004 \cdot stiff\_length \\ &+ 25.5 \cdot stiff\_no\_types - 0.5 \cdot stiff\_no + 0.06 \cdot stiff\_ten\_str - 0.21 \cdot plate\_ten\_str \end{aligned} \quad (15)$$

The equation 15 above has no correlation between the new components and therefore the predictors due to the pre-processing that was performed. On the final test set, the R^2 value is 0.789 and the RMSE value is 234. This is a linear model. However, it will produce more accurate and robust results than equation 14. Likewise the number of optimal components in the model is 4 (components 1,6,7,10). In the training set that underwent 10 fold cross-validations; this had the highest R^2 value.

5.3 SVM regression model

The tuning parameters are sigma = 0.05333; epsilon = 0.0034; C = 10.00. Sigma is the kernel width, epsilon is the width in the epsilon loss function and C is the cost parameter explained earlier. Through 10 fold cross validation it is determined that nine principal components yield the best model with the training data. Once the tuning parameters are applied, it is then possible to plug them into the final equation and verify the model on the final test set. This took one minute to perform. The RMSE value is 42.5 and the R^2 value is 0.993. This is nearly a perfect model. The equation can be represented as follows (Eq. 16):

$$y = -0.190 + \sum_{i=1}^{160} \alpha_i \exp\left(-0.05 \sum_{j=1}^{9} (x_{ij} - x_j^{new})^2\right) \tag{16}$$

The x_{ij} are the set of 160 tuples with the 9 components subtracted from the new interim product whose cost is to be predicted. The α_i changes from 1 to n with n representing the number of tuples in the training set, which are 160. The values for the α_i's are in the table below (See Table 2).

Table 2 Alpha values of the optimized SVM Regression model

```
> svmFit@b
[1] -0.1900191
> alpha<- matrix(0, 160,1)
> alpha[svmFit@alphaindex] <- svmFit@alpha
> c(alpha)
  [1] -10.000000000  -0.143726612  -8.283217018  -2.218607756  10.000000000   0.216592588   4.934494357
  [8]   5.197525610  -3.133760824  -8.068667596 -10.000000000  -6.540373765   0.600006569   8.614781897
 [15]   6.079041879  -1.085714102  -0.219226273   9.953414172   1.642469231  -0.775477624  -1.706159978
 [22]  -6.624775301  -6.097446113  -8.749317929  -4.069564320   1.182343386  -1.614084833 -10.000000000
 [29]   1.205210991  10.000000000  10.000000000   0.907669419  10.000000000 -10.000000000 -10.000000000
 [36]   4.768417329  -2.802109502  10.000000000  -2.541403734  -5.155271808  -7.241892778  -6.681166183
 [43]   6.668769599  -0.044962151  10.000000000   2.933499335  -3.840453748  -3.101346946  -2.338851722
 [50]   1.537358253   0.000000000  -6.663218475   9.984005616 -10.000000000  -6.417863454  -1.082686233
 [57]   0.550539491   1.588783708  -1.590811367  -5.723053716  -8.466198525  10.000000000  -6.751279314
 [64]  -4.727868586   7.755800472  -3.682262108  -5.613827981   4.336073720   7.943569363  -5.770486053
 [71]   2.745074130   3.559804832  -0.104341174  -2.013217285   6.064799830 -10.000000000  10.000000000
 [78]   0.385351560  -6.475048287  -7.209423421   4.609204103   2.935852821   0.000000000   0.000000000
 [85]   0.000000000  -4.874892059  -0.008479049  -2.511276538  -1.684747461   0.000000000   8.511444659
 [92]   7.541177230   7.618430072   7.064559897 -10.000000000  -3.107616997 -10.000000000  -1.626501331
 [99]  -7.309006074 -10.000000000   4.397200917  -3.125951409   0.077645443   0.000000000  -2.396781012
[106]  -1.516968876  -1.307371121  -0.823102560   1.724484808   0.140417610   0.219682013  -1.523226398
[113]   1.874066529   0.058224515  -2.326626998   2.910228494  10.000000000  -8.765961589  -2.092490376
[120]   7.553593293   0.000000000   0.000000000   1.815777040  -0.342686264  10.000000000   4.157627538
[127]   3.196304403 -10.000000000  -6.311877582  10.000000000   2.689798328  10.000000000   7.378210588
[134]  10.000000000  -3.436151395   7.983085500   0.167853016  10.000000000  -9.970987262  -1.862163405
[141]   0.000000000  10.000000000   8.065506112   0.434280036  -7.491313423 -10.000000000  -2.021776215
[148]   3.684428235  -0.249304630   7.322768174   4.696347886  -1.825810618 -10.000000000   5.197176722
[155]   6.386732423 -10.000000000  10.000000000   3.709196819  -1.594467329   0.000000000
```

5.4 Artificial Neural Network model

The ANN model was optimized with 10 fold cross-validation to find the tuning parameters and the number of principal components to be used. This takes the computer 2.5 minutes to perform. The optimized parameters are 10 hidden units, a weight decay of 1, and 10 principal components in the ANN model. This is indicated by the blue oval in the middle of the plot (See Figure 4). Figure 4 aims to represent a three dimensional plot on a two-dimensional surface. The vertical axis in the left side of the plot (y-axis) represents the number of hidden units (ranging from 0 to 10). The blue oval indicates a value 10 hidden units. The horizontal axis (x-axis) shows the log of the weight decay with a base of 10. The oval marks a weight decay of 1 because 10 raised to the 0 is 1. Finally, the third dimension, shown by the vertical axis to the right of the plot (z-axis), represents the number of principal components which is evident through the degree of darkness, ranging from 10 to 17. Therefore, the blue oval is marking a cell which indicates that the darkest shading is 10 which is the number of principal components. The RMSE value is 40.63 and the R^2 value is 0.994, which is better than the SVM regression model. The equation for the ANN uses the coefficients developed from the internal calculations in R-studio program which yields

Equations 17 and 18 below. Making even slight changes by either decreasing or increasing the number of hidden units will result in a non-optimal model which is less accurate and has lower R^2 values. The significance of using a training set and then a test set is part of the data mining methodology. Likewise, the pre-processing to standardize the data along with a decreased number of principal components means that the model is robust in addition to being accurate, which is a new approach in comparison to methods done in previous works, where the pre-processing is not performed.

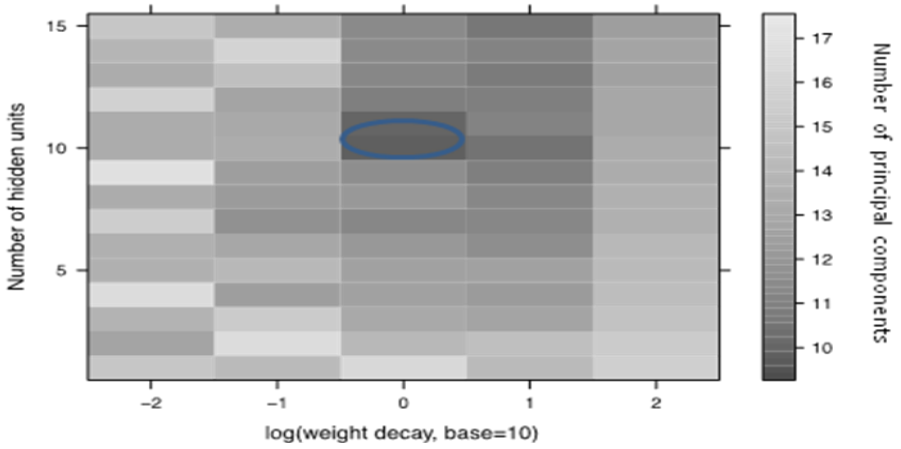

Fig. 4 Three-dimensional plot showing number of hidden units and the weight decay and RMSE values of PCs in the ANN model.

The h's represent the hidden units where the coefficients are also taken from the table above. See equation 17 below for a representation of the equation for the first hidden unit.

$$\begin{aligned} costs = {} & 66 - 66h_1 + 73h_2 - 184h_3 + 284h_4 + 252h_5 - 238h_6 + 164h_7 + 184h_8 \\ & + 175h_9 - 99PC_1 + 19PC_2 - 88PC_3 - 186PC_4 + 4PC_5 + 2.5PC_6 + 41PC_7 \\ & + 77PC_8 + 41PC_9 + 19PC_{10} \end{aligned} \tag{17}$$

Please note that this is a sigmoidal function.

$$\begin{aligned} h = g(10 & - 0.3PC_1 - 1.3PC_2 - 8.2PC_3 - 1.0PC_4 + 9.5PC_5 - 1.2PC_6 \\ & + 9PC_7 + 5.4PC_8 - 6.3PC_9 + 5.9PC_{10}) \end{aligned} \tag{18}$$

5.5 Comparison of different models

The plots of predicted costs vs. actual costs for each of the four methods explained in this paper were made. There is a similar amount of discrepancy between the predicted and the actual costs of the OLS and PCR models (See Figures 5a and 5b). The Figures 5a and 5b show considerable discrepancy since there are many data points which are not along the straight line. The discrepancies between the predicted and actual costs for the SVM regression and ANN models are considerably less than the OLS and PCR models, which is in compliance to the similar R^2 and RMSE values (See Figures 5c and 5d). Many more data points are on or close to the straight line. Since the costs of the multitude of different interim products for a ship are multi-dimensional and non-linear, the OLS and PCR models can never be accurate for predicting costs as are the SVM regression and ANN models. It is important to note that the closer the R^2 value is to one, the more accurate the model. Therefore, for the ANN, the value is 0.994 and for the SVM regression model it is 0.993 which is virtually the same. Likewise the RMSE values are also similar 40.63 for the ANN and 42.5 for the SVM regression model. The ANN model is the most appropriate because it is more accurate (See Table 3).

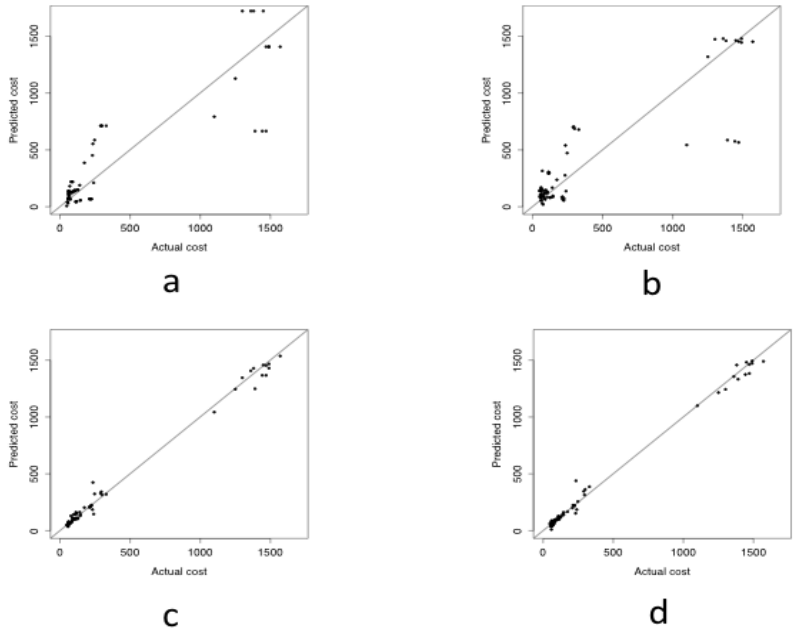

Fig. 5 Actual Costs vs. Predicted Costs for the a) OLS b) PCR model c) ANN model d) SVM regression model

Table 3 Comparison of values between the different models

	OLS	PCR	ANN	SVR
Time to tune 1000 times (sec)	5	6	180	90
Selected no. of predictors / PCs	6	4	10	9
RMSE	230.6	234	40.63	42.5
R^2	0.795	0.789	0.994	0.993
Absolute error	145.9	136.47	23.47	25.85
Mean of true costs	364.3	364.3	364.3	364.3
Mean predicted costs	380.6	369.5	366.4	365.9
Std. dev. of true costs	508.7	508.7	508.7	508.7
Std. dev. predicted costs	495.1	457.4	506.2	496.1
New panel cost	1702	1442	1316	1342
Actual panel cost	1360	1360	1360	1360
Difference from actual cost	342	82	-44	-18

5.6 Validation of models with a new interim product

The purpose of creating the above models is to predict the costs of new interim products that are different from the tuples of the training and the test sets. Therefore this was done using a panel with the following predictors or characteristics: mass= 23968 kg; plate thickness = 16 mm; panel length = 12000 mm; panel width = 11050 mm; stiffener height = 370 mm;

stiffener thickness = 13 mm; stiffener length = 132600 mm (this represents the total length of all of the stiffeners on the panel); stiffener number of types = 1; stiffener number = 12; stiffener tensile strength = 235 N/mm^2; plate tensile strength = 235 N/mm^2. The predicted costs can be seen in the table below. Please note that the actual cost is 1360 Euros.

For this specific interim product example, the SVM regression model predicts a cost slightly closer to the actual panel cost. Then follows the ANN model, which is also very close. However, since the R^2 value of ANN is still slightly greater, on average it will produce more accurate values as the number of interim products analyzed increase. The PCR model is third in accuracy and finally the ordinary least squares (OLS) model predicts a value farthest from the actual cost (See Table 3).

6. Conclusions

The aim of developing an accurate model to predict the assembly costs of interim products is successful. Regardless of whether that interim product is a larger panel or one of three different micro-panels, by creating a table with multiple tuples, it is possible to train a model that will accurately predict costs of different types of interim products that have still not been designed or produced at the shipyard. It is important to analyse the tuples with all predictors and first determine whether it is necessary to do pre-processing and if so what pre-processing. It can be expected that there will be some correlation between predictors of the same tuple. Likewise, the magnitudes of multiple predictors have different units and therefore different scales such as mm and kg and N/mm^2. Therefore it can be expected that in most cases analysis of the data will result in the necessity to perform pre-processing of scaling, centring and PCA.

Upon pre-processing, it is necessary to determine which model to use. In the case of cost prediction of shipyard interim products, since a linear model is not accurate and due to the large number of predictors it is necessary to have both a robust and accurate model that can be used by the shipyard. This leads to the development of a tailor made methodology of reaping the benefits of different methods but determining which one is best for a specific shipyard's interim products cost prediction. The SVM regression takes less time to fine tune and there is only one local minimum. Since the SVM regression requires inversion of the matrix of tuples it can handle data set sizes of 10000's of tuples, but computation becomes difficult for much larger sample sizes. Therefore, the ANN, which does not do matrix inversion, will be the best model for the job. It can be expected that initially a shipyard will have tuples that represent a typical cross section of a standard ship. However, since the bow and sterns of a vessel have different configurations than the typical cross section it is realistic to assume that tuples will eventually approach 10000 and more. Therefore, the ANN model will be more realistic. When tuples are in the range of fewer than 10000 tuples then the SVM regression model is faster and simpler.

7. Recommendations for future work

For future work, the authors recommend this tailor made methodology to be tried on the interim products of another shipyard or another industry. One the data is collected, the predictors or characteristics of the interim products will be different as will the costs. However the same data mining methods of pre-processing and decreasing the number of components will be crucial. The hypothesis is that either the SVM regression or ANN models will be the most accurate. However, in the case of industries where there is a more linear relationship between the cost and the characteristics, it is possible that the OLS may be accurate if there is more of a linear relationship.

REFERENCES

[1] Agard, B., Kusiak, A., Data-mining-based methodology for the design of product families, Data-mining-based methodology for the design of product families." International Journal of Production Research 42.15 (2004): 2955-2969.

[2] Bhuddharaju, P, Laskar S, Samuel G.R. (2007) Robust well-cost estimation using a support vector machine model, Society of Petroleum Engineers – Digital Energy Conference and Exhibition – Bytes and Barrels: An Energy Renaissance, Houston, Texas, pp. 28-33

[3] Caprace JD, Rigo P, Warnotte R, Le Viol S (2006) An analytical cost assessment module for the detailed design stage fuzzy metric for assessing the producibility of straightening in early design, Conference on Computer Applications and Information Technology in the Maritime Industries (COMPIT), Delft, Netherlands, pp. 1-11

[4] Choudhary AK, Harding JA, Tiwari MK (2009) Data mining in manufacturing: A review based on the kind of knowledge, J Intell Manuf (20)**:**501-521, doi:10.1007/s10845-008-0145-x

[5] Das, SK, Samui P, Sabat AK (2012) Prediction of field hydraulic conductivity of clay liners using an artificial neural network and support vector machine, Int J Geomech (12): 606-611, doi:10.1061/(ASCE)GM.1943-5622.0000129.

[6] Fafandjel N, Zamarin A, Hadjina M (2010) Shipyard cost structure optimisation model related to product type, International Journal of Production Research 48(5):1479-1491, doi:10.1080/00207540802609665

[7] Feng CJ, Yu Z, Kusiak A (2006) Selection and validation of predictive regression and neural network models based on designed experiments, IIE Transactions 38(1):13-23, doi:10.1080/07408170500346378

[8] Han J, Kamber M (2006) Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufmann Publishers, ISBN 978-0123814791

[9] Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference and prediction, Springer, New York

[10] Jin Y, Curran R, Burke R, Welch B (2012) An integration methodology for automated recurring cost prediction using digital manufacturing technology, International Journal of Computer Integrated Manufacturing (25): 326-339, doi:10.1080/0951192X.2011.570171

[11] Kim SH, Lee CM (1997) Nonlinear Prediction of Manufacturing Systems through Explicit and Implicit Data Mining, Computers Ind Engng 33(3-4)**:**461-464

[12] Kolich D, Fafandjel N, Zamarin A (2012) Lean manufacturing methodology for shipyards, Brodogradnja – Journal of Naval Architecture and Shipbuilding Industry 63(1):18-29

[13] Kolich D, Storch RL, Fafandjel N (2012) Value stream mapping methodology for pre-assembly steel processes in shipbuilding, Proceedings of the International Conference on Innovative Technologies, Rijeka, Croatia, pp*:*365-368, ISBN 978-953-6326-77-8.

[14] Kolich D, Yao YL, Fafandjel N, Hadjina M (2014) Value stream mapping micropanel assembly with clustering to improve flow in a shipyard, Proceedings of the International Conference on Innovative Technologies, IN-TECH, 10–13 Sept 2014, Leiria, Portugal, ISBN: 978-953-6326-88-4, pp 85-88

[15] Kuhn M, Johnson K (2013) Applied Predictive Modelling, Springer, New York ISBN 978-1-4614-6848

[16] Lamorski K, Pachepsky Y, Stawinski C (2008) Using Support Vector Machines to Develop Pedotransfer Functions for Water Retention of Soils in Poland, Soil Science Society of America Journal 72 (5): 1243-1247.

[17] Mansfield SD, Kang K, Iliadis L, Tachos S, Avramidis S (2011) Predicting the strength of populus spp clones using artificial neural networks and ε-regression support vector machines (ε-rSVM) Holzforschung 65:855-863, doi:10.1515/HF.2011.107.

[18] Mere, JB, Marcos A, Gonzalez JA, Rubio V (2004) Estimation of mechanical properties of steel strip in hot dip galvanising lines, Ironmaking and Steelmaking 31**:**45-50

[19] Mikovsky P, Matouek K, Kouba Z (2002) Data pre-processing support for data mining, Proceedings of the IEEE International Conference on Systems, Man and Cybernetics 5:51-54 ISSN:02243627

[20] Nan R, Yan-Xin H, Bing-Jie H (2013) Research on complex product WBS process simulation and prediction, Information Technology Journal (12) 6: 3380-3384, doi:10.3923/itj.2013.3880.3884

[21] Osma, A (2014) An investigation on the stress-strain relationship of cold-rolled steel sheets used in the automotive industry, Proc IMech Part D: J Automobile Engineering, 228(5): 565-579

[22] Razavi AR, Gill H, Ahlfeldt H, Shahsavar N (2005) A data pre-processing method to increase efficiency and accuracy in data mining, Artificial Intelligence in Medicine. 10th Conference on Artificial Intelligence in Medicine, AIME, pp 434-443, ISBN-10:3 540 27831 1

[23] Ren Y, Ding Y, Zhou S (2006) A data mining approach to study the significance of nonlinearity in multistation assembly processes, IIE Transactions (38)12:1069-1083, doi:10.1080/07408170600735538.

[24] Shi G (2008) Superiorities of support vector machine in fracture prediction and gassiness evaluation, Petroleum Exploration and Development (35)6:588-594

[25] Storch RL, Hammon CP, Bunch HM, Moore RC (1995) Ship Production, SNAME, New Jersey.

[26] Stockton DJ, Khalil RA, Mukhongo LM (2013) Cost model development using virtual manufacturing and data mining: part I – comparison of data mining algorithms Int J Adv Manuf Technol, 66:1389-1396 doi:10.1007/s00170-012-4416-5

[27] Taghavifar H, Mardani A (2014) A comparative trend in forecasting ability of artificial neural networks and regressive support vector machine methodologies for energy dissipation modelling of off-road vehicles, Energy, 66:569-576 http://dx.doi.org/10.1016/j.energy.2014.01.022

[28] Trumbule JC, Dougherty JJ, Deschamps L, Ewing R, Greenwell CR, Lamb T (1999) Product oriented design and construction (PODAC) cost model – an update, Paper Presented at the Ship Production Symposium, Arlington, Virginia

[29] Yan X, Zhang C, Zhang S (2003) Towards databases mining: pre-processing collected data, Applied Artificial Intelligence, 17:545-561.

[30] Yuan B, Wang XZ, Morris T (2000) Software analyser design using data mining technology for toxicity prediction of aqueous effluents, Waste Management, 20:677-686

Submitted: 07.12.2015.
Accepted: 28.01.2016

Damir Kolich, dkolic@riteh.hr, Nikša Fafandjel
University of Rijeka-Faculty of Engineering, Vukovarska 58, 51000 Rijeka, Croatia
Y. Lawrence Yao
Columbia University, 220 Mudd building, MC 473, 500 West 120th Street, New York, NY 10027 USA