sciendo

CIVIL AND ENVIRONMENTAL ENGINEERING REPORTS

E-ISSN 2450-8594

CEER 2020; 30 (3): 0048-0067
DOI: 10.2478/ceer-2020-0033
Original Research Article

DEVELOPMENT OF COST ESTIMATION MODELS BASED ON ANN ENSEMBLES AND THE SVM METHOD

Michał JUSZCZYK[1],
Cracow University of Technology, Faculty of Civil Engineering

Abstract

Cost estimation, as one of the key processes in construction projects, provides the basis for a number of project-related decisions. This paper presents some results of studies on the application of artificial intelligence and machine learning in cost estimation. The research developed three original models based either on ensembles of neural networks or on support vector machines for the cost prediction of the floor structural frames of buildings. According to the criteria of general metrics (*RMSE*, *MAPE*), the three models demonstrate similar predictive performance. *MAPE* values computed for the training and testing of the three developed models range between 5% and 6%. The accuracy of cost predictions given by the three developed models is acceptable for the cost estimates of the floor structural frames of buildings in the early design stage of the construction project. Analysis of error distribution revealed a degree of superiority for the model based on support vector machines.

Keywords: construction cost estimation, cost modelling, ensembles of neural networks, support vector machine

1. INTRODUCTION

Cost estimation is a key process for any construction project. The objective of the process is to deliver forecasts of construction costs on the basis of information available on successive stages of projects. The accuracy of the forecasts has a significant impact on project success as a number of decisions are made on the basis of

[1] Corresponding author: Cracow University of Technology, Faculty of Civil Engineering, Warszawska 24, 31-155 Kraków, Polska, mjuszczyk@L7.pk.edu.pl, +48 12 628 30 90

cost analyses. This paper presents some results of studies on the applicability of artificial intelligence and machine learning-based methods for the process of estimating construction costs. Alternative models are introduced which are based on either ensembles of artificial neural networks (ANN) or on the support vector machine method (SVM).

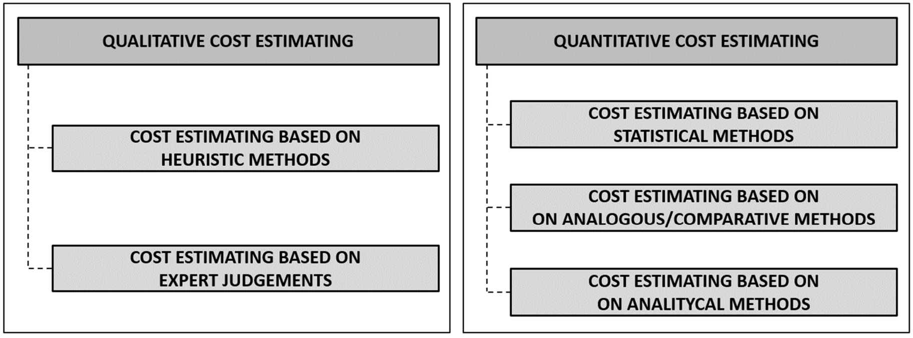

QUALITATIVE COST ESTIMATING
COST ESTIMATING BASED ON HEURISTIC METHODS
COST ESTIMATING BASED ON EXPERT JUDGEMENTS
QUANTITATIVE COST ESTIMATING
COST ESTIMATING BASED ON STATISTICAL METHODS
COST ESTIMATING BASED ON ON ANALOGOUS/COMPARATIVE METHODS
COST ESTIMATING BASED ON ON ANALITYCAL METHODS

Fig. 1. Classification of cost estimation methods with regard to methodology

The proposal of classification of cost estimating methods is presented in Figure 1 (compare [9, 13, 24, 31]). According to this classification, cost estimation based on statistical methods belongs to a broad class of quantitative methods. Both ANN and SVM are rooted in advanced statistics, their application for cost estimation relies on regression analysis. The fundamental assumption is that there exists a relationship between cost and a set of cost predictors. The former is considered a dependent variable while the latter are independent variables. The set of the known values of the variables allows the development of cost estimation models which are supposed to map the relationship. Specifically, nonparametric cost estimation is based on fitting an unknown, implicit function to the data representing cost and cost predictors – there is no functional relationship assumed *a priori*. (For more theoretical details about nonparametric estimating, one can refer to [33]).

It is important to note that there exist a variety of construction cost estimation problems, which differ with regard to cost analysis. Therefore, the variables of models that are built for certain problems must be individually selected for each problem. The development of advanced statistical methods, especially those that are built on artificial intelligence and machine learning, along with the increasing data storage and processing capacities, has resulted in the exploration of cost estimation based on ANN or SVM for construction. Some examples of the use of ANN in the field are as follows:

- modelling costs of various facilities or structures such as highways [35], road tunnels [27], sports fields [18] and buildings [6, 29];
- forecasting of construction site overhead costs [7, 25].

Examples of the use of SVM for construction cost estimation problems are:

- prediction of construction project costs at completion [4];

- forecasting construction project cost [21];
- prediction of bridge construction costs [16].

Some of the works present and compare both ANN and SVM methods for:

- estimating costs of school buildings [20];
- modelling cost and schedule success for construction projects [34].

The use of both of the discussed tools is obviously not limited to cost estimation problems in construction. Some examples of applications of ensembles of ANN in broadly defined engineering problems are: modelling of air pollution [3], forecasting of concrete compressive strength for high-performance concretes [8], prediction of buildings' electricity load levels [12], and the analysis of labour efficiency in construction works [17]. With regard to SVM, the use of this method is reported *inter alia* for: aiding contractors' prequalification decision making processes [1], predicting heavy machinery performance in earthworks [22], analysis of noise pollution in special protection areas [23], forecasting of building energy consumption [37].

The aim of this work was to develop and compare models where either ANN ensembles or the SVM method were implemented to support fast cost estimates of floor structural frames of residential buildings. The applicability of these methods is discussed on the basis of the obtained results. This work continues and extends previous research [14, 15].

2. METHODOLOGY

In the course of the research, several cost estimation models based on nonparametric statistical methods were developed. The models were designed to provide predictions of the construction costs of the floor structural frames of buildings. The introduced models were based either on ANN ensembles or on the SVM method. Both the former and the latter were implemented for the problem as supervised learning models that allowed regression analysis and the implicit realisation of the relationships between costs and cost predictors.

The theory, fundamentals and details for both methods that were omitted for the sake of brevity in this paper can be found in the literature for ANN see, for example, [2, 11, 26, 32] and for SVM see, for example, [5, 10, 30, 36].

The basic assumption for the use of ANN ensembles is to combine a set of trained ANN and to use this set to approximate a true regression function instead of using a single ANN. Various kinds of ANN or ANN trained to different local minima might be incorporated into the ensemble (compare [2]). Such an approach brings a degree of reduction to the model's error compared to the single network-based models. Moreover, it is useful for practical implementations in problems for which the number of training data samples is not large.

The rationale for the use of SVM is the method's capability to deal with high dimensional data. SVM enables finding a global solution for a given task, it also works well on relatively small sets of training data. The use of both of these methods makes it possible to take into account several cost predictors (describing variables) and modelling relationships that bind these cost predictors with the construction costs of floor structural frames of buildings.

Three models were developed in the course of research:

- an ANN ensemble model based on a generalised averaging approach (later referred to as ANN $\mathrm{ENS}_{\mathrm{GA}}$),
- an ANN ensemble model based on a stacked generalisation approach (later referred to as ANN $\mathrm{ENS}_{\mathrm{SG}}$),
- a model based on SVM regression (later referred to as $\mathrm{SVM}_{\mathrm{REG}}$).

The following subsections present assumptions for the development of models and the concise presentation of data used for the purposes of supervised learning and testing.

2.1. Assumptions for the development of the models

Let y be a dependent variable (expected model's output) and let $\boldsymbol{x}$ be a vector of the independent variables (model's input) – specifically:

- y – construction cost of the building's floor structural frame,
- $\boldsymbol{x}$ – information and characteristics of the building, structural and material solutions, basic measures of quantities.

Consequently, the nonparametric cost estimating model is expected to implement input-output mapping: $\boldsymbol{x} \rightarrow y$. Variables (both y and $\boldsymbol{x}$) that were used in the course of the models' development process are explained in detail in section 2.2.

If y denotes the expected model output, i.e. the real-life values of the dependant variable (real-life values of construction costs of the building's floor structural frames), then let $\hat{y}$ be the values predicted by a certain model. The error for p–th data sample is then e^p:

$$e^p = y^p - \hat{y}^p \tag{2.1}$$

Consequently, the assumed general metrics of the models' cost prediction performance are: Pearson's correlation coefficient (R), root mean squared error ($RMSE$), mean absolute percentage error ($MAPE$):

$$R = \frac{cov(y,\hat{y})}{\sigma_y \sigma_{\hat{y}}} \tag{2.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_p (e^p)^2} \tag{2.3}$$

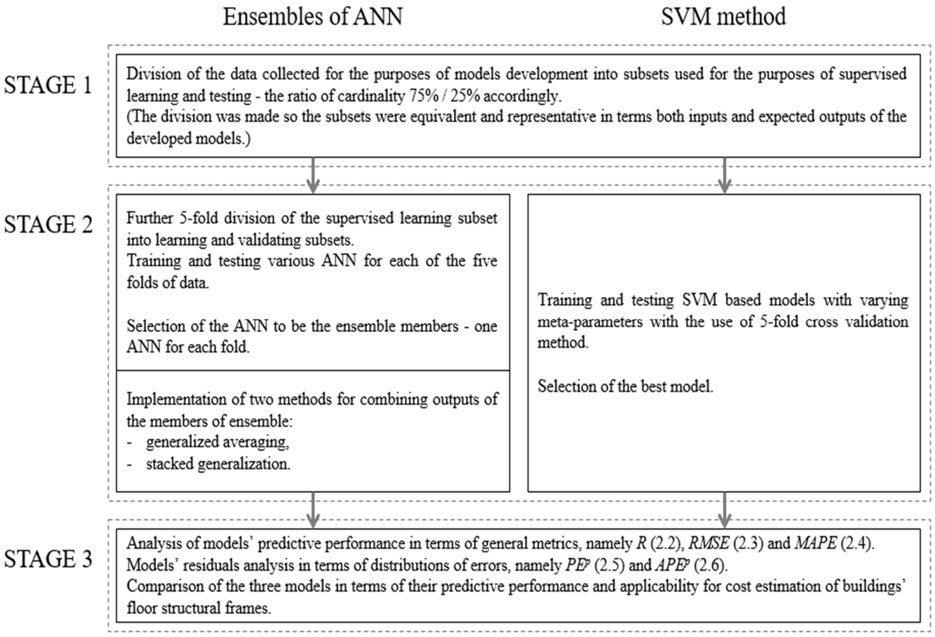
$$MAPE = \frac{100\%}{n}\sum_p \left|\frac{e^p}{y^p}\right| \tag{2.4}$$

where: *cov(y,ŷ)* – covariance between y and $\hat{y}$, σ_y – standard deviation for y, $\sigma_{\hat{y}}$ – standard deviation for $\hat{y}$, p – index of a data sample belonging to one of the subsets (for models based on ANN to L or V, for models based on SVM to L or T).
Additionally, the measures for residuals' analyses were:

$$PE^p = \left(\frac{e^p}{y^p}\right)100\% \tag{2.5}$$

$$APE^p = \left|\frac{e^p}{y^p}\right|100\% \tag{2.6}$$

Successive stages of the models' development and the assessment of their performance are presented schematically in Figure 2. For the two models based on ANN ensembles, the members of the ensembles were chosen from a set of trained ANN. The ANN differ from each other in their structures and employed activation functions – each of the selected members of the ensemble were trained for one of the five folds of testing and validating data (the details can be found in the scheme).



Ensembles of ANN

SVM method

STAGE 1

Division of the data collected for the purposes of models development into subsets used for the purposes of supervised learning and testing - the ratio of cardinality 75% / 25% accordingly.
(The division was made so the subsets were equivalent and representative in terms both inputs and expected outputs of the developed models.)

STAGE 2

Further 5-fold division of the supervised learning subset into learning and validating subsets.
Training and testing various ANN for each of the five folds of data.

Selection of the ANN to be the ensemble members - one ANN for each fold.

Implementation of two methods for combining outputs of the members of ensemble:
- generalized averaging,
- stacked generalization.

Training and testing SVM based models with varying meta-parameters with the use of 5-fold cross validation method.

Selection of the best model.

STAGE 3

Analysis of models' predictive performance in terms of general metrics, namely *R* (2.2), *RMSE* (2.3) and *MAPE* (2.4).
Models' residuals analysis in terms of distributions of errors, namely PE^p (2.5) and APE^p (2.6).
Comparison of the three models in terms of their predictive performance and applicability for cost estimation of buildings' floor structural frames.

Fig. 2. Process of the models development and assessment

In the case of the generalised averaging approach (assumed for model ANN ENS$_{GA}$) the assumption is that members of an ensemble are linearly combined (compare [2, 11]) so that the output of an ensemble $\hat{y}$ is computed as the weighted average of ANN members outputs $\hat{y}_k$, so that:

$$\hat{y} = \sum\nolimits_{k} \hat{y}_k \alpha_k \tag{2.7}$$

where α_k stands for weight coefficients assigned to k-th ANN member of an ensemble, $k = 1, \ldots, K$, and the following conditions are fulfilled:

$$\sum\nolimits_{k} \alpha_k = 1 \tag{2.8}$$

$$\alpha_k = \frac{\sum_{l=1}^{K} (M^{-1})_{kl}}{\sum_{h=1}^{K} \sum_{l=1}^{K} (M^{-1})_{hl}} \tag{2.9}$$

where M is an error correlation matrix of errors produced by the members of an ensemble. In eq. (2.9), the member networks are marked by indexes k, h and l for clarity.

In the case of the stacked generalisation approach (assumed for model ANN ENS$_{SG}$), the general assumption is that the members of an ensemble are level-0 models for which outputs are combined with the use of the level-1 model. A scheme of this approach is presented in Figure 3.

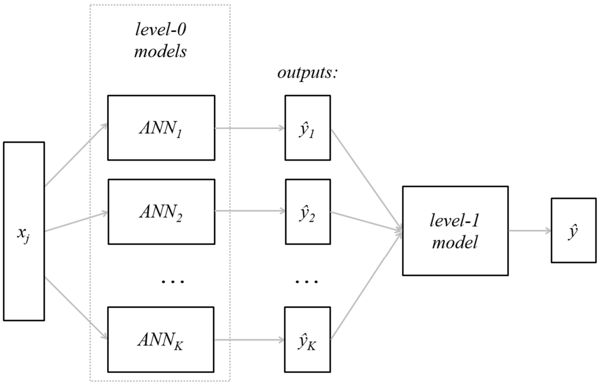


Fig. 3. Scheme of stacked generalisation

For this study, it was assumed that the linear regression model would be used as the level-1 model. Thus, the prediction can be formally given as:

$$\hat{y} = b_0 + b_1\hat{y}_1 + b_2\hat{y}_2 + \cdots + b_K\hat{y}_K \tag{2.10}$$

where b_0, b_1, b_2 , …, b_K are the structural parameters of the level-1 model.

The SVM method application for the given regression problem was based on an approximation of the assumed mapping as a linear regression hyperplane. The hyperplane was computed with the use of an SVM method after the nonlinear transformation of the input training data ***x*** to the high dimensional linear feature space with the use of a kernel trick and the application of a nonlinear kernel function.
The aim was to find an approximation hyperplane which minimises the generalisation error:

$$\frac{1}{2}||\omega|| + C\sum(\xi + \xi^{*}) \rightarrow min \tag{2.11}$$

Where ω is the sought for hyperplane's parameter, C stands for the complexity parameter of a model, ξ and ξ^{*} are slack variables introduced to make the method less prone to noise and outliers. Slack variables are computed for each training data sample, in particular: ξ above, and ξ^{*} below the ε parameter of Vapnik's loss function (ε determines borders within which the approximated hyperplane must lie – for details see, for example, [30, 36]) so that the constraints for eq. (2.11) are:

$$\begin{cases} y^{p} - < x^{p}, \omega > \leq \varepsilon + \xi \\ -y^{p} + < x^{p}, \omega > \leq \varepsilon + \xi^{*} \\ \xi \geq 0;\ \xi^{*} \geq 0 \end{cases} \tag{2.12}$$

The optimisation problem is solved with the use of Lagrange multipliers. Support vectors are the data points that correspond to non-zero multipliers for the optimal solution. Thus, the support vectors influence the position of the approximated hyperplane. Moreover, the use of the chosen kernel function *K* and scalar products *K(x,x')* (the so-called kernel trick) is also implemented. Finally, the prediction can be formally given as:

$$\hat{y} = \sum_{i}(\alpha_i - \alpha_i^{*})\,K(x, x') + \widehat{\omega}_0 \tag{2.13}$$

Where α and α^{*} are the Lagrange multipliers for the optimal solution.

2.2. Variables used in the course of the development of the models

In accordance with the general assumptions for the development of the models, the dependent variable can be explained as follows:

- total construction cost of the floor structural frame of a building – the variable took numerical values corresponding to costs given in thousands of PLN excluding VAT and was denoted as *y*.

Information brought to the models by independent variables (cost predictors) was related to: building size, its location, geometrical measures of the building's floor and structural members of its frame, basic material characteristics and some quality measures of construction works. In particular, the independent variables included such data as:

- building height, (with regard to classes present in relevant Polish legal acts) – the variable took one of the three possible nominal values: low, medium-high or high, that were coded as *one-of-n* and denoted as: x_1 for 1, 0, 0, x_2 for 0, 1, 0 and x_3 for 0, 0, 1;
- gross floor area – the variable took numerical values corresponding to the measured surface given in m^2 and denoted as x_4;
- volume of horizontal structural members made of reinforced concrete (including slabs, beams, landings and flights of stairs) – the variable took the numerical values of the measured cubic capacity given in m^3 and denoted as x_5;
- volume of vertical structural members made of reinforced concrete (including walls and columns) – the variable took the numerical values of the measured cubic capacity given in m^3 and denoted as x_6;
- class of concrete – the variable originally took nominal values corresponding to the class of concrete for the structural members that were *pseudo-fuzzy* scaled to values 0.1, 0.5 or 0.9 and denoted as x_7;
- class of formwork execution – the variable originally took nominal values such as class 1, class 2 or class 3 that were *pseudo-fuzzy* scaled to values 0.1, 0.5 or 0.9, respectively and denoted as x_8;
- volume of vertical masonry structural members (including walls) – the variable took the numerical values of the measured cubic capacity given in m^3 and denoted as x_9;
- class of masonry works execution – the variable originally took nominal values such as: class A or class B that were *pseudo-fuzzy* scaled to values 0.9 or 0.1 and denoted as x_{10};
- location of building – the variable originally took nominal values corresponding to the relevant voivodship of Poland, that were *pseudo-fuzzy* scaled to values 0.1, 0.3, 0.5, 0.7 or 0.9 and denoted as x_{11}.

In the case of variables x_7, x_8, x_{10}, x_{11} *pseudo-fuzzy* scaling into numerical values was made with regard to the association of the nominal values of these four variables with the costs of construction works. The increase of numerical values was related to growth of the costs.

Table 1 presents descriptive statistics for the variables that took numerical values.

Table 1. Descriptive statistics for variables that took numerical values

Symbol	Mean	Standard Deviation	1^{st} quartile	2^{nd} quartile	3^{rd} quartile	4^{th} quartile
y	194.45	72.72	148.85	184.80	232.75	418.40
x_4	384.29	160.82	207.60	316.80	482.15	884.50
x_5	82.00	37.19	58.75	69.30	95.85	197.80
x_6	28.87	17.34	19.95	27.15	34.18	111.47
x_9	73.91	28.15	47.20	78.20	94.95	139.70

Table 2 presents the frequencies of values for variables $x_1 – x_3$ coded as *one-of-n*. The frequencies of values that were taken by variables coded with the use of a pseudo fuzzy scale are presented in Table 3.

The total number of samples that were used for model development (both for the purposes of training and testing models) was 162. The details of the data division into training and testing subsets are explained in the scheme depicted in Figure 2.

Table 2. Frequencies of values for building height class coded as *one-of-n*

Symbol	1, 0, 0	0, 1, 0	0, 0, 1
x_1	25.77%	-	-
x_2	-	44.79%	-
x_3	-	-	29.45%

Table 3. Frequencies of variables values coded with the use of *pseudo-fuzzy* scale

Symbol	0.1	0.3	0.5	0.7	0.9
x_7	11.04%	-	46.01%	-	42.94%
x_8	31.29%	-	39.88%	-	28.83%
x_{10}	51.15%	-	-	-	47.85%
x_{11}	23.31%	19.63%	19.02%	25.77%	12.27%

3. RESULTS

For each of the five folds of learning and validating subsets number of various ANN of a multilayer perceptron type were trained (see the scheme in Figure 1). The ANN differed in terms of their structure – the number of neurons in the networks' hidden layers varied between 2 to 8. Moreover, various activation functions (namely: exponential (EXP), logistic (LOG), hyperbolic tangent (TANH) and linear (LIN) - in both the hidden and the output layer) were considered. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm was used for the purposes of supervised learning.

For each of the 5 folds of training the data number of the candidate ANN were investigated. Assessment of their performance enabled the selection of 5 ANN (one ANN for each of the folds) to be the members of the ensemble-based model.

Details regarding structures, activation functions and *RMSE* values computed for the ensemble members are presented in Table 4.

The criteria of selection reflected expectations of equivalence of learning, validating and testing errors and a high correlation of y and $\hat{y}$ that is expected, and also predicted the total construction costs of a floor structural frame of a building for learning, validating and testing subsets. In Table 4, one can see that the *RMSE* values are relatively close for each of the subsets and for all of the networks. For all of the selected ANN, R > 0.960 for each of the subsets.

Table 4. Details of ANN that were selected to be the members of the ensemble

k-th fold	ANN structure	Activation functions hidden layer / output layer	Training algorithm	$RMSE_L$	$RMSE_V$	$RMSE_T$
1	11_7_1	EXP / LIN	BFGS	16.742	16.369	17.012
2	11_7_1	EXP / LOG		16.348	16.594	16.639
3	11_5_1	EXP / LOG		16.392	16.546	17.230
4	11_8_1	EXP / LIN		16.160	16.294	15.731
5	11_3_1	TANH / LIN		17.198	16.093	16.405

Coefficients α_k for the ANN ENS$_{GA}$ model were computed with the use of eq. (2.9). The values of α_k are given below:

$$\alpha_1 = 0.1760;\ \alpha_2 = 0.2962;\ \alpha_3 = 0.1761;\ \alpha_4 = 0.2344;\ \alpha_5 = 0.1173$$

In the case of ANN ENS$_{SG}$, structural parameters of the level-1 model (see eq. (2.10)) were found with the use of the commonly known linear regression analysis and the least squares method. The parameters are given as follows (standard estimation errors for each of the parameters are given in the brackets below):

$$\underset{(3.555)}{b_0 = -8.9556};\ \underset{(0.0620)}{b_1 = 0.1336};\ \underset{(0.0618)}{b_2 = 0.2920};\ \underset{(0.0644)}{b_3 = 0.0490};\ \underset{(0.0637)}{b_4 = 0.9024};\ \underset{(0.0696)}{b_5 = 0.4808}$$

For both of the ANN ensemble-based models, the outputs, which are the predictions of the construction costs of the building's floor structural frame, were computed with the use of the coefficients α_k for ANN ENS$_{GA}$ and structural parameters b_k for ANN ENS$_{SG}$ given above. The outputs were computed for training and testing subsets of data on the basis of eq. (2.7) and eq. (2.10), respectively.

In the case of the SVM method, a number of models were investigated in order to find the one to implement cost prediction mapping: $\boldsymbol{x} \rightarrow y$. For the investigated models, a radial basis function was assumed as a kernel function:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2) \tag{3.1}$$

In the course of the research, a range of meta-parameters of the models was analysed. To find the model with the best performance, the values of C, ε and γ were sought with the use of grid analysis. The grid was characterised by ranges of values and steps specified for each of the parameters:

- in the case of ε, the values varied between 0.05 and 0.10 (step 0.01);
- in the case of C, the values varied between 1 and 20 (step 1);
- in the case of γ, the values varied between 0.05 and 0.15 (step 0.01).

In the course of the computations, 5-fold cross validation was applied (see scheme in Figure 1). A number of SVM based models were investigated and analysed. Details for the five models with the best performance are presented in Table 5. For the five models, one can see the values of C, ε and γ as well as the number of unbound vectors – uv and bound vectors – bv, parameter ω_0 and cross validation error cv_{err}. It was found that the best performance was obtained for ε = 0.05, this is reflected in Table 5.

Table 5. Five SVM-based models with the best performance

mod.	C	γ	ε	uv	bv	ω_0	cv_{err}	$RMSE_L$	$RMSE_T$
1	20	0.05	0.05	61	29	0.884	0.009	15.684	15.738
2	16	0.06	0.05	62	28	0.783	0.009	15.699	15.770
3	11	0.07	0.05	61	28	0.678	0.009	15.711	16.028
4	9	0.08	0.05	63	27	0.607	0.009	15.711	16.099
5	9	0.09	0.05	62	27	0.559	0.009	15.658	16.057

The selection criteria for SVM was similar to that used for ANN. Equivalence of training and testing errors and a high correlation of y and $\hat{y}$ were expected. In Table 5, one can see *RMSE* values as performance measures. *RMSE* values were relatively close for both the subsets used for supervised learning and those for the testing models. For all of the models presented in the table, R > 0.970 for each of the subsets of data.

Finally, it was decided that model number 1, as the model with lowest *RMSE* values for which R_L=0.976 and R_T=0.978, would be implemented as the core of the SVM$_{\mathrm{REG}}$ for predictions of the construction costs of the building's floor structural frame. The outputs of the model were computed for training and testing subsets of data.

Comparison of the ANN ENS$_{\mathrm{GA}}$, ANN ENS$_{\mathrm{SG}}$ and SVM$_{\mathrm{REG}}$ models' predictive performance of the total construction costs of a building's floor structural frames in terms of general metrics is presented in Table 6. The values of *R*, *RMSE* and *MAPE* are given for the training and testing of models. One can see that the differences between the models with regard to the values of certain general metrics are relatively small, especially where the number of training and testing samples is concerned.

Analysis of the values presented in Table 6 allows to conclude that the general performance metrics are comparable for the three models.

Table 6. Comparison of general performance metrics for the three developed models

Model / perf. metr.	ANN ENS$_{GA}$		ANN ENS$_{SG}$		SVM$_{REG}$	
	TRAIN.	TEST.	TRAIN.	TEST.	TRAIN.	TEST.
R	0.978	0.980	0.978	0.984	0.976	0.978
RMSE	14.616	15.747	12.976	15.984	15.684	15.738
MAPE	5.55%	5.38%	5.22%	6.36%	5.75%	4.92%

Figure 4 depicts a comparison of expected outputs y (denoted in the figures as “exp”, given in ascending order) and corresponding predictions of building’s floor structural frame construction costs $\hat{y}$ by the three developed models. Figure 4a presents results for training, and Figure 4b presents results for testing. From the graphs in Figure 3, one can generalise that predictions follow the expected outputs in similar ways for all three models.

Figures 5, 6 and 7 present distributions of PE^p errors for the three developed models computed for training and testing. The errors were counted in the range widths of 5% with consideration as to whether the values were positive or negative. The shapes of the graphs depicting PE^p errors for training are similar in the case of all of the three models. In the case of the PE^p testing errors, one can see some differences between the models. Considering the graphs for each of the three models individually, especially while comparing training and testing errors, one can see that distributions of training and testing errors appear most similar in the case of the SVM$_{REG}$ model.

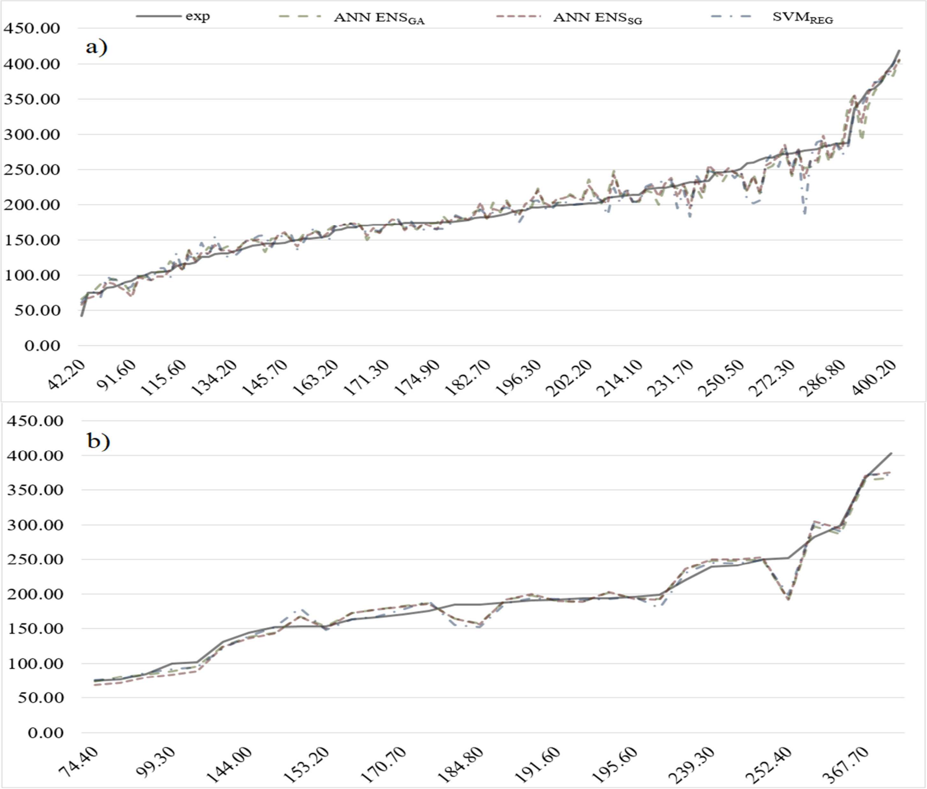


Fig. 4. Comparison of expected outputs and predictions of the three developed models: a) training subset, b) testing subset

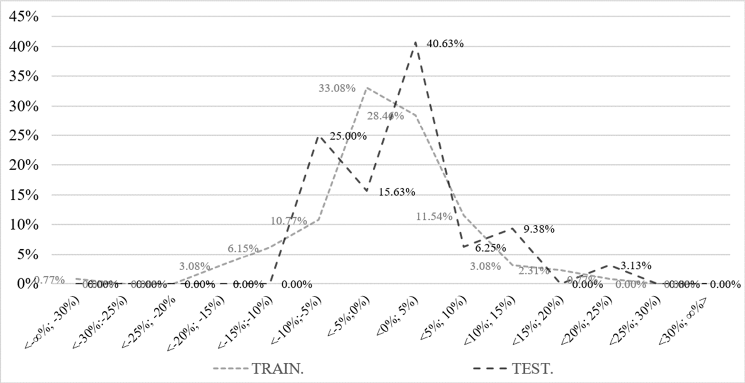


Fig. 5. Distribution of PE^{p} errors - ANN ENS$_{GA}$

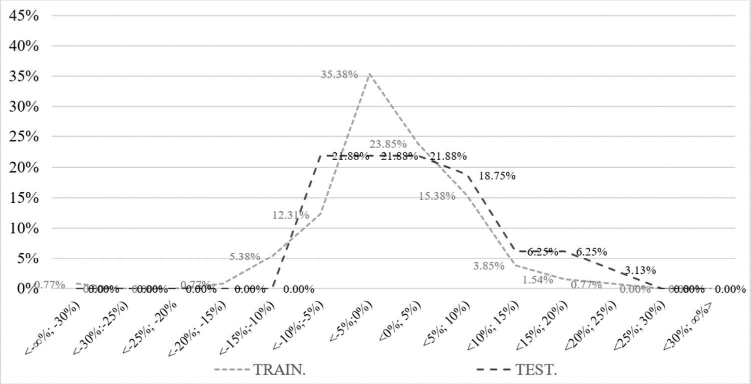


Fig. 6. Distribution of PE^p errors - ANN ENS_{SG}

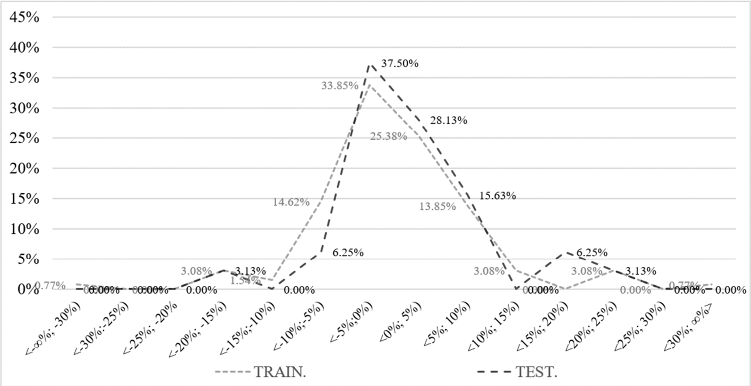


Fig. 7. Distribution of PE^p errors – SVM_{REG}

Table 7 presents the accumulated shares of APE^p errors. The values were counted and accumulated in ranges increasing by 5% with each step. The table shows that in the case of all three developed models, more than 95% of APE^p errors (computed for both teaching and testing samples) were smaller than or, at most, equal to 20%. Differences between the models result from the increase in the number of errors in individual ranges $APE^p \leq 5\%$, $APE^p \leq 10\%$ and $APE^p \leq 15\%$, which can be seen both for training and testing subsets.

Table 7. Cumulative distribution of APE^P values for the three developed models computed for training and testing subsets

Model / APE^P cum. dist.	ANN ENS_{GA}		ANN ENS_{SG}		SVM_{REG}	
	TRAIN.	TEST.	TRAIN.	TEST.	TRAIN.	TEST.
$APE^P \leq 5\%$	61.54%	56.25%	59.23%	43.75%	59.23%	65.63%
$APE^P \leq 10\%$	83.85%	87.50%	86.92%	84.38%	87.69%	87.50%
$APE^P \leq 15\%$	93.08%	96.88%	96.15%	90.63%	92.31%	87.50%
$APE^P \leq 20\%$	98.46%	96.88%	98.46%	96.88%	95.38%	96.88%
$APE^P \leq 25\%$	99.23%	100.00%	99.23%	100.00%	98.46%	100.00%
$APE^P \leq 30\%$	99.23%	100.00%	99.23%	100.00%	98.46%	100.00%
$APE^P > 30\%$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

In terms of the general performance metrics, all three of the introduced models offer satisfactory prediction performance. With regard to the accuracy of cost predictions, the models fulfil the expectations for estimates provided in the early stage of the construction project as most of the percentage errors of cost predictions fall in the range of<-20%;+20%>.

4. DISCUSSION

The development of cost-estimation models based on ANN and SVM as tools rooted in artificial intelligence and machine learning is an up-to-date area of research and publications in the field of construction management. In the authors' opinion, one of the main reasons for this is the need for the introduction of new methods that are alternatives to the traditional approach and capable of aiding cost estimates, especially in the early phase of construction projects. The application of ANN or SVM brings the following benefits:

- the results of cost estimates are based on the relationship of cost with multiple describing variables related to analysed characteristics of objects, quantity measures and technical parameters;
- there is no need to assume *a priori* functional relationships between the cost and describing variables for regression analysis;
- cost estimates are based on the collected information (training patterns) which form the basis for automated training processes and gaining knowledge;
- the developed models provide cost estimates in a very short time – it is also possible to analyse many variants that differ from each other in the values of describing variables.

Success in the development of models based on artificial intelligence and machine learning that offer satisfactory performance of cost predictions depends on overcoming a major obstacle – the collection of data and information necessary for supervised training process is a challenging task in itself. From the authors previous experience and research (see: [14, 15, 16, 18, 25]) it follows that for construction cost estimation problems, it is most likely to collect datasets that include a moderate amount of data. However, this matter may be counterbalanced through the use of ensembles of ANN or SVM. The tools work well for small or moderate datasets even if one must solve high dimensional problems (both of the mentioned methods apply to construction cost estimation problems).

The use of both ensembles of ANN and SVM for the investigated problem of cost estimation of elements of buildings – namely floor structural frames – brought the expected benefits. Models are developed for multidimensional problems – there are twelve describing variables that provide information about the floor structural frames. Moreover, there was no need to assume functional relationships between the described variable, that is cost of construction works and the describing variables. The models enable quick cost estimation of the specified building element.

The development of the two models based on ANN ensemble needed more computational effort when compared to similar models based on single ANN (see [14]), this is reflected by the two step procedure (see Fig. 2),. On the other hand, the ensemble approach and the implementation of five combined ANN in the two introduced models (ANN ENS_{GA}, ANN ENS_{SG}) resulted in a synergy effect and the compensation of cost estimation errors obtained for the ANN acting in isolation.

The third of the introduced models, which was based on the SVM method, required determination of the kernel function and ranges of values for meta-parameters. Several candidate models were trained with the use of cross-validation for tuning meta-parameters, from which one was finally selected to be implemented for the cost estimation problem (SVM_{REG}).

For models based on ensembles of ANN as well as for the model based on SVM, the correlation of expected and predicted values of the construction costs of buildings' floor structural frames is high (both for training and testing). General performance metrics are comparable for the three models and lead to the conclusion that their predictive capabilities are satisfactory. In particular, values of *MAPE* errors (see Table 6) confirm the applicability of the developed models in the investigated cost estimation problem. Analysis of the distribution of PE^p and APE^p errors leads to two main conclusions: firstly, the models are predestined to cost estimates in the early stage of the construction project; secondly, the model based on the SVM method appears to be somewhat better than the two models based on ANN ensembles due to its more stable results of training and testing (see Figures 6-7).

Most of the models presented in the literature aiming to support cost estimates in construction projects and based on artificial intelligence or machine learning are focused on various types of construction objects as a whole. The models presented herein are developed to aid cost estimates of certain elements of construction objects –

parts of buildings that require completion of a complex of construction works. Thus, the introduced models may be used for quick variant cost analyses – the considered variants may differ in the values of parameters specific for members of floor structural frames representing results of certain types of construction works.

5. CONCLUDING REMARKS

The research presented herein allowed investigation of the applicability of artificial intelligence and machine learning tools in estimating the costs of buildings' floor structural frames. The research resulted in the development of the three models capable of aiding cost estimates. The developed models were based on:
- ensemble of 5 ANN and generalised averaging approach – ANN ENS_{GA};
- ensemble of 5 ANN and stacked generalisation approach – ANN ENS_{SG};
- SVM method – SVM_{REG}.

All of the three models offer comparable performance in cost prediction in terms of general metrics (especially *RMSE* and *MAPE* errors). The obtained accuracy of cost estimates of the structural frames of building floors is acceptable for the early design stage of a construction project. Analysis of the distribution of training and testing errors for each model showed some superiority for the model based on support vector machines.

ADDITIONAL INFORMATION

Computations for ANN and SVM based models were made with the use of the TIBCO Statistica™ software suite.

REFERENCES

1. Attar, A, M, Khanzadi, M, Dabirian, S and Kalhor, E 2013. Forecasting contractor's deviation from the client objectives in prequalification model using support vector regression, *International Journal of Project Management* **31(6)**, 924-936.
2. Bishop C, M 1995. *Neural networks for pattern recognition.* Oxford University Press.
3. Bougoudis, I, Iliadis, L and Papaleonidas, A 2014. Fuzzy inference ANN ensembles for air pollutants modeling in a major urban area: the case of Athens. In: Mladenov, V, Jayne, C, Iliadis, L, (eds) Engineering Applications of Neural Networks. EANN 2014. *Communications in Computer and Information Science* **459**, Cham: Springer, 1-14.

4. Cheng, M, Y and Hoang, N, D 2014. Interval estimation of construction cost at completion using least squares support vector machine. *Journal of Civil Engineering and Management* **20(2)**, 223-236.
5. Cristianini, N and Shawe-Taylor, J 2000. *An Introduction to Support Vector Machines (and Other Kernel-based Learning Methods)*. Cambridge: Cambridge University Press.
6. El-Sawalhi, N, I and Shehatto, O 2014. A Neural Network Model for Building Construction Projects Cost Estimating. *Journal of Construction Engineering and Project Management* **4(4)**, 9–16.
7. El-Sawy, I, Y, Hosny, H, E and Razek, M, A 2011. A Neural Network Model for Construction Projects Site Overhead Cost Estimating in Egypt. *International Journal of Computer Science Issues* **8(3)**, 273-283.
8. Erdal, H, I, Karakurt, O and Namli, E 2013. High performance concrete compressive strength forecasting using ensemble models based on discrete wavelet transform. *Engineering Applications of Artificial Intelligence* **26(4)**, 1246-1254.
9. Foussier, P, M, M 2006. *From Product Description to Cost: A Practical Approach, vol.1: The Parametric Approach.* Berlin: Springer.
10. Gunn, S, R 1997. *Support Vector Machines for Classification and Regression. Technical Report.* Southampton: University of Southampton, Image Speech and Intelligent Systems Research Group.
11. Haykin, S 1998. *Neural Networks: A Comprehensive Foundation*, Prentice Hall.
12. Jetcheva, J, G, Majidpour, M and Chen, W, P 2014. Neural network model ensembles for building-level electricity load forecasts. *Energy and Buildings* **84**, 214-223.
13. Juszczyk, M 2017. The challenges of nonparametric cost estimation of construction works with the use of artificial intelligence tools. *Procedia engineering* **196**, 415-422.
14. Juszczyk, M 2018. Implementation of the ANNs ensembles in macro-BIM cost estimates of buildings' floor structural frames. *AIP Conference Proceedings* **1946(1)**, 020014
15. Juszczyk, M 2019. Cost Estimates of Buildings' Floor Structural Frames with the Use of Support Vector Regression. *IOP Conference Series: Earth and Environmental Science* **222(1)**, 012007.
16. Juszczyk, M 2020. On the Search of Models for Early Cost Estimate of Bridges: An SVM-Based Approach. *Buildings* **10(1), 2**, 1-17.
17. Juszczyk, M 2020. Analysis of labour efficiency supported by the ensembles of neural networks on the example of steel reinforcement works. *Archives of Civil Engineering* **66(1)**, 97-111.
18. Juszczyk, M, Leśniak, A and Zima, K 2018. ANN Based Approach for Estimation of Construction Costs of Sports Fields. *Complexity* **2018**, 1-11.

19. Kasprowicz, T 2007. Inżynieria przedsięwzięć budowlanych in Kapliński O, (ed.) *Metody i modele badań w inżynierii przedsięwzięć budowlanych.* Warszawa: Polska Akademia Nauk, Komitet Inżynierii Lądowej i Wodnej, 35-78.
20. Kim, GH, Shin, JM, Kim, S and Shin, Y 2013. Comparison of School Building Construction Costs Estimation Methods Using Regression Analysis, Neural Network and Support Vector Machine. *Journal of Building Construction and Planning Research* **2013.1**, 1-7.
21. Kong, F, Wu, X and Cai, L 2008. *Application of RS-SVM in construction project cost forecasting.* WiCOM'08 - 4th International Conference on Wireless Communications, Networking and Mobile Computing, 1.
22. Mahdevari, S, Shahriar, K, Yagiz, S and Shirazi, M, A 2014. A support vector regression model for predicting tunnel boring machine penetration rates, *International Journal of Rock Mechanics and Mining Sciences* **72**, 214-229.
23. Mrówczyńska, M, Sztubecka, M, Skiba, M, Bazan-Krzywoszańska, A and Bejga, P, 2019. The Use of Artificial Intelligence as a Tool Supporting Sustainable Development Local Policy, *Sustainability* **11(15)**, 1-17.
24. Layer, A, Brine, ET, Van Houten, F, Kals, H and Haasis, S 2002. Recent and future trends in cost estimation. *International Journal of Computer Integrated Manufacturing* **15(6)**, 499–510.
25. Leśniak, A and Juszczyk, M 2018. Prediction of site overhead costs with the use of artificial neural network based model. *Archives of Civil and Mechanical Engineering* **18(3)**, 973-982.
26. Osowski, S 2004. *Sieci neuronowe do przetwarzania informacji.* Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej.
27. Petroutsatou, K, Georgopoulos, E, Lambropoulos, S and Pantouvakis, J, P 2012. Early Cost Estimating of Road Tunnel Construction Using Neural Networks. *Journal of Construction Engineering and Management* **138(6)**, 679–687.
28. Potts, K 2008. *Construction cost management: learning from case studies.* Taylor&Francis.
29. Roxas, CLC and Ongpeng J, M, C 2014, *An Artificial Neural Network Approach to Structural Cost Estimation of Building Projects in the Philippines.* De La Salle University Research Congress, Manila:DLSU, 1-7.
30. Smola, A, J and Schölkopf, B 2004. A tutorial on support vector regression. *Statistics and computing* **14.3**, 199-222.
31. Stewart, RD and Wyskida, R, M 1987. *Cost Estimator's Reference Manual.* New York: Wiley.
32. Tadeusiewicz, R 1993. *Sieci neuronowe.* Warszawa: Akademicka Oficyna Wydawnicza.
33. Tsybakov, AB 2008. *Introduction to nonparametric estimation.* Paris: Springer.
34. Wang, YR, Yu, CY and Chan, HH 2012. Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines

classification models, *International Journal of Project Management* **30(4)**, 470-478.

35. Wilmot, CG and Mei, B 2008. Neural network modeling of highway construction costs. *Journal of Construction Engineering and Management* **131(7)**, 765-771.
36. Vapnik, V 2013. *The Nature of Statistical Learning Theory*. New York: Springer.
37. Zhang, F, Deb, C, Lee, SE, Yang, J and Shah, KW 2016. Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique, *Energy and Buildings* **126**, 94-103.

Editor received the manuscript: 09.06.2020