Primljen / Received: 15.7.2020.
Ispravljen / Corrected: 31.1.2021.
Prihvaćen / Accepted: 3.6.2021.
Dostupno online / Available online: 10.9.2021.

A simple formulation for early-stage cost estimation of building construction projects

Authors:

Esra Dobrucali, PhD. CE
Sakarya University, Turkey
Faculty of Engineering
Departement of Civil Engineering
eeken@sakarya.edu.tr
Corresponding author

Assist.Prof. **Ismail Hakki Demir**, PhD. CE
Sakarya University, Turkey
Faculty of Art Design and Architecture
Departement of Architecture
idemir@sakarya.edu.tr

Research Paper

Esra Dobrucali, Ismail Hakki Demir

A simple formulation for early-stage cost estimation of building construction projects

This study is aimed at improving a formula that enables easy, correct, and fast estimation of an Early-Stage Cost of Buildings (ESCE). This formula, enabling estimation of ESCE, was developed by the authors based on artificial neural networks and gene expression programming. A quantity survey was conducted for a hundred construction projects, and a data set was created. This data set was analysed with many Artificial Neural Networks to determine the variables that affect ESCE. An algorithm configuration was made with Gene Expression Programming, and the ESCE formula was created using this algorithm configuration. This formula estimates ESCE with satisfactory precision. The use of the proposed formula in the early-stage building cost calculations is important not only for faster and easier cost calculation but also to prevent any differences that may arise due to the individual making the calculations

Key words:
early-stage cost of buildings, artificial neural network, gene expression programming, construction project, cost estimation

Prethodno priopćenje

Esra Dobrucali, Ismail Hakki Demir

Jednostavna formula za ranu procjenu troškova na građevinskim projektima

U okviru ovog istraživanja poboljšana je formula koja omogućuje jednostavnu, točnu i brzu procjenu troškova u ranim fazama građevinskih projekata (ESCE). Spomenutu formulu za procjenu ESCE-a razvili su autori na temelju umjetnih neuronskih mreža i evolucijskog programiranja gena. Kvantitativna analiza provedena je na stotinu građevinskih projekata, te je izrađen odgovarajući niz podataka. Taj niz podataka analiziran je pomoću većeg broja umjetnih neuronskih mreža kako bi se odredile varijable koje utječu na procjenu ESCE-a. Konfiguracija algoritma provedena je pomoću evolucijskog programiranja gena, te je na temelju te konfiguracije izrađena formula ESCE. Ta formula omogućuje dovoljno preciznu procjenu ESCE-a. Primjena predložene formule za određivanje troškova u ranoj fazi projekta omogućuje brže i jednostavnije izračunavanje troškova, ali isto tako sprječava pojavu bilo kakvih razlika do kojih bi moglo doći zbog individualnog pristupa proračunu.

Ključne riječi:
troškovi u ranim fazama građenja, umjetna neuronska mreža, evolucijsko programiranje gena, građevinski projekt, procjena troškova

1. Introduction

On construction projects, investors wish to accurately estimate necessary financing so as to arrange the budgets of their investments, and to make sure that the project is profitable. In this process, cost calculation is an important step that should be taken into consideration by all parties, including project owners and contractors [1, 2]. For this purpose, prices of similar finished works, archived data of experienced contractors, and price determinations from other institutions such as market research or related occupational chambers and universities, are used to obtain correct data for the Early-Stage Cost Estimation of Buildings Construction Projects (ESCE).

In the case of construction projects, and especially in public investments tendered with limited resources, the most accurate estimation of construction costs is one of the most important issues in terms of construction management at the pre-tendering stage. Public institutions acting as employers wish to obtain accurate results in cost estimation calculations in the shortest period of time so as to obtain the necessary financing, and to arrange the budgets of their investments, whereas the bidders / potential contractors wish to calculate their cost and profits accurately, and give the best offer in a competitive environment. During feasibility studies, at the preliminary design stage, or where the bidding period is limited, approximate costs may need to be calculated as soon as practicable. While bidding for a tender, if a detailed cost analysis is not possible due to time constraints, the control of the early cost calculation with a simple ESCE method may be quite convenient for competing companies and public utility companies. An early-stage project cost calculation can also vary from one individual to another. For these reasons, it is obvious that there is a need for methods that would enable accurate estimation of ESCE, both in practice and theory.

Since 1998, researchers have conducted many modelling studies in order to estimate costs of buildings at an earlier stage. Elhag and Boussabaine [3] studied the cost estimation of construction projects and developed two artificial neural network models using 14 variables. Lowe et al. [4] conducted a study using multiple regression analysis to predict construction costs. In his study, Hwang [5] proposed two distinct regression models for price estimation on construction projects. Cheng et al. [6] integrated artificial intelligence techniques in their study for the conceptual cost estimation of construction projects in Taiwan. Kim et al. [7] (12) conducted a study to measure performance of artificial neural networks, support vector techniques, and regression analysis methods in the early determination of construction costs. Besides, Bostancioglu [8], Gunaydın and Dogan [9], Akinbingol and Gultekin [10], Dogan et al. [11], Nan et al. [12], Sonmez [13], Arafa and Alqedra [14], Kuruoglu et al. [15], Cho et al. [16], Latief et al. [17], El-Sawalhi and Shehatto [18], Bayram et al. [19], Coloma et al. [20], Dimitrijević et al. [21] conducted modelling studies in order to estimate costs of buildings at an earlier stage. The literature review is detailed in Table 1.

Table 1. Literature reviews.

Year	Author	Method
1998	Elhag and Boussabaine	Artificial neural networks
2004	Gunaydin and Dogan	Regression analyses
2009	Hwang	Regression analyses
2006	Dogan et al.	Artificial intelligence techniques
2006	Lowe	Regression analyses
2010	Cheng et al.	Artificial intelligence techniques
2011	Sönmez	Artificial neural networks - bootstrap
2011	Arafa and Alqedra	Artificial neural networks
2013	Kim et al.	Regression analyses - artificial neural networks - support vector techniques
2013	Cho et al.	Regression analyses - artificial neural networks
2013	Latief et al.	Regression analyses - artificial neural networks
2014	El-Sawalhi and Shehatto	Artificial neural networks
2016	Bayram et al.	-
2018	Dimitrijević et al.	Regression analyses

AI has been known to be a fine modelling technique in various fields including the specialty areas within civil engineering [22]. Artificial Neural Network (ANN), Genetic Algorithm (GA), Gene Expression Programming (GEP) and simulation have recently been used, either combined or separately, in various fields of civil engineering. According to Table 1, researchers have generally used Artificial Intelligence Techniques (AI) and Regression Analyses (RA) for the estimation of Building Cost at Early Stage and investigated the efficiency of these techniques to predict ESCE. Studies in the literature confirm validity of AI in the assessment of overhead construction costs. To the best of the authors' knowledge, no formula that estimates ESCE has so far been presented in the literature. Therefore, the first aim of this study is to propose a formula to calculate ESCE in a rapid, easy, and accurate manner. Another aim is to show that ESCE models of satisfactory precision can in fact be created with AI.

For this purpose, a new formula is proposed as a result of ANN-GEP integration model with a specific number of parameters selected from architectural and static projects. The proposed ESCE formula can help investors and bidders not only to calculate costs faster and easier, but the formula

also prevents any differences in ESCE calculations that may arise due to the individual approaches.

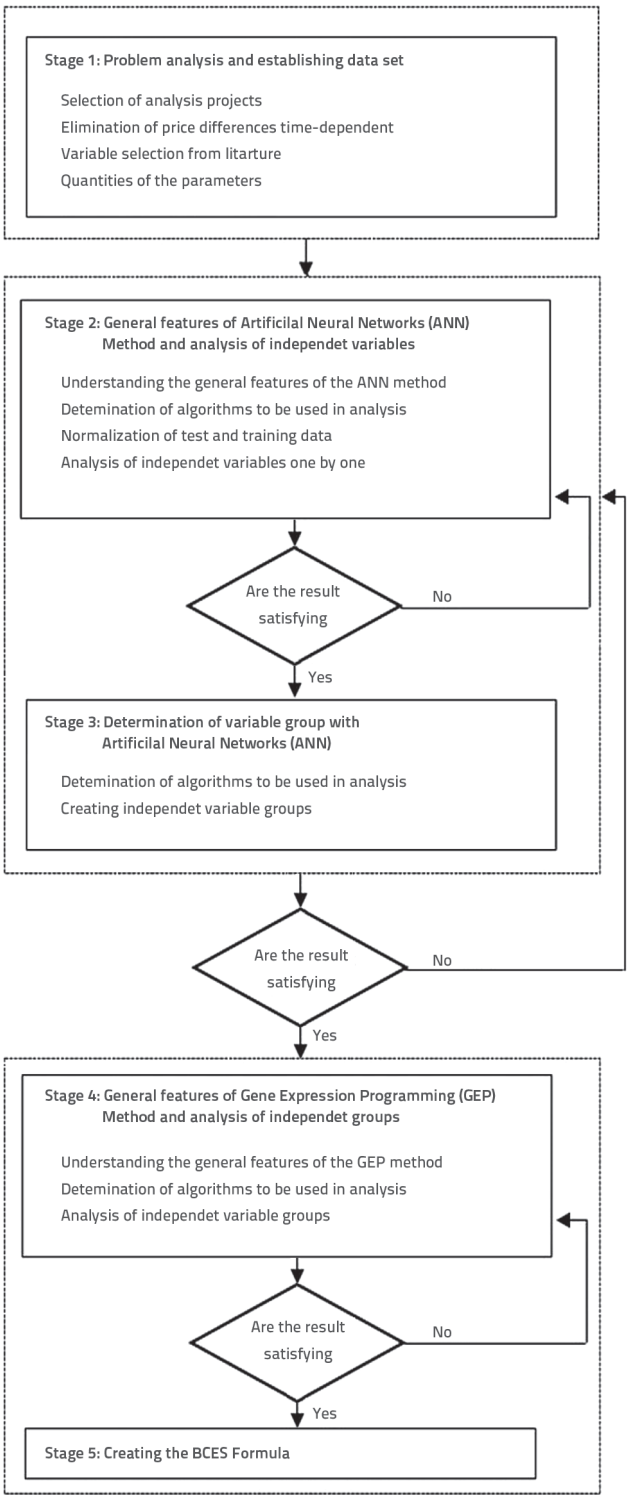


Figure 1. Research methodology (Research methodology adapted based on Leśniak and Juszczyk [23] methodology)

2. Methodology

The authors' assumption was to improve a formula that enables easy, correct and fast estimation of building cost. The methodology shown in Figure 1 was applied to obtain this formula. In the mentioned methodology, after conducting extensive literature research, 18 independent variables were evaluated and grouped according to their ability to predict ESCE using the ANN method. Then, the algorithm of the group with the highest ability to predict ESCE was designed using GEP, and the GEP model was created. Finally, a formula was created in which ESCE was calculated correctly. Stage 1 is described in Section 3. The detailed information about the formula and other stages are given in Section 4.

3. Problem analysis and establishment of data set

Early-Stage Cost of Buildings is one of the most critical components of the employer's and contractor's budget accounts on a construction project. Employers wish to obtain accurate results in cost estimation calculations in the shortest period of time in order to arrange the budget of their investments whereas the contractors wish to calculate their profits accurately. The traditional estimation of building costs based on quantity survey calculation for the whole project is also accurate but time consuming. The cost calculation at an early stage of projects can vary from one individual to another. At the preliminary design stage or where the bidding period is limited, building costs need to be calculated in an expedited manner. While bidding for a tender, if a detailed cost analysis is not possible due to time constraints, the control of the early cost calculation with a simple ESCE method may prove quite convenient for bidding companies and public utility companies. For these reasons, it is apparent that there is a need for methods to estimate ESCE accurately both in theory and practice. To meet this need, this study proposes an alternative formula that enables easy, correct and fast estimation of the Early-Stage Cost of Buildings.

Prices of similar finished works, archived data of experienced contractors, and price determinations from other institutions such as market research or related occupational chambers and universities, are used to obtain correct data about the Early-Stage Cost of Buildings (ESCE). Therefore, the formula of ESCE was designed using a hundred public reinforced concrete construction projects tendered in different regions and cities of Turkey between 2011 and 2016. The total floor area of projects used in this study was selected as a variable and it ranged between 141 m^2 to 7947 m^2.

The cost of projects becomes comparable on a common basis, after elimination of time-dependent price differences. ESCE of the projects had to be adjusted to take into account changing economic conditions by selecting a reference year, because construction projects under study were realized in various time periods between 2011 and 2016.

Table 2. Selection of independent variables from literature

Variable \ References	[13]	[7]	[9]	[14]	[18]	[6]
Year		x				
Function					x	
Social area	x					
Ground floor area / building area			x			
Floor area / building area			x			
Building area / usage area						
Building area	x	x				x
Ground floor area				x		
Floor area				x	x	
Building cost index	x					
City cost index	x					
Floor height		x				
Number of floors	x	x	x	x	x	
Maintenance area / total area	x					
Parking area / total area	x					
Housing unit total gross construction area	x					
Construction zone	x	x	x			x
Demolition amount	x					
Purification	x					
Wooden frame	x					
Steel frame	x					
Concrete frame	x					
Iron and concrete frame	x					
Masonry building	x					
Wood outer coating	x					
Plastic outer coating	x					
Stone exterior wall cladding	x					
Plaster exterior wall cladding	x					
Number of elevator stops	x					
Project duration	x					
Number of consoles in the building			x			
Soil structure-topography						x
Basic type			x	x	x	
Construction type			x			

For this reason, the real ESCE in the analysis set were adjusted using the "construction cost index and rate [24]" of change by the reference year 2016 (Equation 1). Additionally, according to the Central Bank of the Republic of Turkey 2019 exchange rate data, 1 TL was assumed to be $ 5.7.

$$ESCE = \frac{CCI_x}{CCI_i} x\, Real\, ESCE \quad (1)$$

Here *ESCE* refers to the Adjusted Early-Stage Cost of Buildings, CCI_x refers to the Construction Cost Index for the desired year, CCI_i stands for the Construction Cost Index for 2016, and Real ESCE refers to ESCE calculated by employers/public institutions. Studies conducted by Cheng et al. [6], Sonmez [13], Arafa and Alqedra [14], Kim et al. [7], El-Sawalhi, Shehatto [18] and Gunaydin and Dogan [9] on "early stage building costs" were taken as a basis for the selection of independent variables that will be subject to quantity survey. Eighteen parameters selected in this context are shown in Table 2. Quantities of the parameters specified for each of the mentioned projects were calculated with exact individually performed measurements.

4. Analyses

4.1. General features of Artificial Neural Networks (ANN) method and analysis of independent variables

Artificial Neural Networks (ANN) are one of the artificial intelligence techniques that function on the basis of current examples of learning abilities of the human brain. The ANN is supplying a contemporary, accurate and matchless solution based on artificial intelligence [25]. ANN, which is not algorithmic and which can carry out parallel actions efficiently, can solve complicated and non-linear problems in a serial and convenient manner [26]. The selection of the right network in ANN is the most important stage for learning the network. Network topology, addition function, activation function and learning strategy are the elements which distinguish one network model from other network models [27]. The most fundamental part of ANN is the nerve cell and it is also called the "Processor". The processor is comprised of one or more entries, weights depending on the entries, integration function (total connections), transfer (activation) function, and a single outcome.

The integration function provides net entries by processing between the inputs coming to the cell and weights. The most commonly used function, among the functions that can assume the form of addition, multiplication, maximum, minimum, majority etc., is the addition function [27]. The activation function, which is also called the "Learning Curve", is used to create the outcome value [28]. Several types of activation functions are being used, such as sigmoid, linear, step, sinus, hyperbolic tangent, etc. These functions are selected by the user who creates the network [27]. Sigmoid and hyperbolic tangent functions are the most widely used transfer functions [28].

Depending on the number of layers, the following two types of ANN can be differentiated: single-layer networks or multi-

Table 3. Results of ANN analysis where independent variables are used as a single output

Independent variable name	Number of neurons in hidden layer	BP			SCG		
		R^2	SSE		R^2	SSE	
			Training	Testing		Training	Testing
Type	7	0.4619*	2.1450	0.2476	0.4619*	2.1450	0.2476
Floor area	7	0.8724	0.6523	0.0618	0.8872	0.5806	0.0727
Floor height	7	0.7053	1.3706	0.2404	0.7600	1.1518	0.1669
Number of floors	7	0.6795	1.4705	0.0931	0.6805*	1.4641	0.0900
Duration	7	0.7393	1.2369	0.2303	0.7496	1.9548	0.2478
Ground class	7	0.2298	2.5830	0.4906	0.2299*	2.5826	0.4905
Foundation type	7	0.1394	2.6738	0.4979	0.1394*	2.6738	0.4979
Number of basements	7	0.4522	2.1692	0.2782	0.4522*	2.1692	0.2782
Elevator	7	0.6023*	1.7376	0.0911	0.6023*	1.7376	0.0911
Number of vertical bearers	7	0.7592	1.1566	0.1277	0.8757	0.6356	0.1283
Earthquake zone	7	0.1260	2.6835	0.5293	0.1260*	2.6835	0.5293
Building height	7	0.8228	0839	0.0807	0.8489	0.7621	0.0786
Building importance coefficient	7	0.4112	2.2659	0.3630	0.4114*	2.2652	0.3633
Vertical bearer area	7	0.8601	0.7146	0.1015	0.8946	0.5447	0.0941
Total internal wall area	7	0.8578	0.7249	0.0917	0.8901	0.5662	0.0956
Total exterior wall area	7	0.9485	0.2754	0.2122	0.9560	0.2346	0.2240
Wet area	7	0.9312	0.3642	0.0500	0.9414	0.3101	0.0528
Total indoor area	7	0.9460	0.2880	0.0135	0.9616	0.2054	0.0144

* The training discontinued before iteration reached 10000 as the error curve reached the minimum slope.
BP - Back Propagation, SCG - Scaled Conjugate Gradient

layer networks [29]. A single layer ANN is comprised of entry and exit layers whereas a multi-layer ANN is comprised of three separate layers. These layers are the input layer, hidden layers, and output layer. Multi-layer ANN, which can provide 95 % results for engineering problems, are the models that are nowadays most frequently used [27].

For the ANN analyses, the ESCE, which is the dependent variable, and independent variables were applied to Equation (2), and linear normalization procedure was performed. The objective of this procedure is to facilitate learning and prevent errors [30].

$$y_i = \frac{y}{y_{max}} \tag{2}$$

Here y_i refers to the result of the normalization, y refers to the value that will be normalized, and y_{max} denotes the largest value of the variable.

ANN analyses were done with written codes using the Matlab R2018a program. In the ANN model, the activation (transfer) function was selected as Sigmoid (logsig), which gives positive values between [0-1], whereas the integration function was selected as summation function, which is most commonly used in the literature. Learning with a trainer (consulted) was selected as the learning method. Back propagation (BP) algorithm is the most frequently used algorithm for multi-layer feed forward networks [31]. BP was used as the learning algorithm as it can easily be proven and it is preferred by the studies within the literature review; Scaled Conjugate Gradient (SCG) was used as it yields successful results in problems where BP is used.

4.2. Determination of variable groups with Artificial Neural Networks (ANN)

The important stage in the ANN is to find a way-out optimized network in conformity with the numbers of neurons available based on the hidden layer for the estimation coefficient (R^2). During the first stage of the artificial neural network analyses, the ESCE estimation coefficient (R^2) was determined for the situation where each of the eighteen independent variables was a dependent variable. In these analyses, independent variables make up the input layer individually, whereas the output layer is made of the ESCE. The number of neurons in the hidden layer that gives the best solution was determined to be seven. The number of iterations was found by trial method during each training. It was observed that each variable has a different hidden layer number. Table 1 provides the ANN analysis results for the best hidden layer value of each independent variable. In the analyses, it was adopted that the maximum number of iterations is 10000. However, as the reduction in the slope ratio of the training data would increase the

error margin of the testing data, the network was stopped when the error curve for the training data reached the slope.

Table 4. R^2 values of testing data relating to best variables for ESCE determination

Independent variable name	Training	Testing
Total indoor area	0.9616	0.9603
Total exterior wall area	0.9560	0.5108
Wet area	0.9414	0.9033
Vertical bearer area	0.8946	0.7535
Total internal wall area	0.8901	0.7977
Floor area	0.8872	0.8132
Number of vertical bearer	0.8757	0.7102

Table 5. Independent variables selected for ESCE determination

Independent variable name	Condensation
Total indoor area	y1
Wet area	y2
Total exterior wall area	y3
Total internal wall area	y4
Vertical bearer area	y5
Floor area	y6
Number of vertical bearers	y7

As can be seen in Table 3, the variables with the highest R^2 value in the estimation of ESCE are wet areas and total indoor construction areas. During the first stage of the ANN analyses, independent variables were selected among the variables with training results equal to 0,85 and above, as provided in Table 3. They were then calculated with SCG algorithm. The corresponding values are shown in Table 4. Table 4 shows that the R^2 value in the total exterior wall area is 0,96 for the training data and 0.51 for the testing data. As the R^2 values between the testing and training data were different than expected, when independent variables were being named, the total exterior wall area variable was named as y3 and the wet area variable was named as y2. These are shown in Table 5.

When determining independent variable groups, the ACG1 was first created with total indoor area and wet area (y1 and y2) which are the two variables with the highest R^2 value. Afterwards, five variables other than these two variables were added to ACG1 one by one, and five new groups (ACG2-ACG6) were created. The group with the best analysis result was selected among these five new groups. The other four variables, which were not used, were added to the selected group and new ESCE groups were formed. This process was continued until ACG16 with seven variables was created. The creation of the groups is summarized in Table 6.

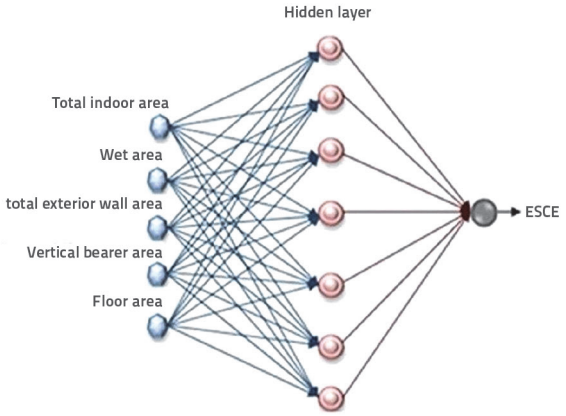


Figure 2. Structure of ANN for ACG 12

Table 6. ANN analysis results for independent variable groups (ACG) in determining ESCE

Independent variable name	Number of neurons in hidden layer	Iteration number	R^2 (Training)	SSE	
				Training	Testing
ACG 1 (y1-y2)	7	400	0.9733	0.1441	0.0137
	7	600	0.9742	0.1392	0.0123
	7	800	0.9751	0.1346	0.0128
ACG 2 (y1-y2-y3)	7	1200	0.9781	0.1184	0.0270
	7	1500	0.9785	0.1161	0.0154
	7	1800	0.9792	0.1126	0.0262
ACG 3 (y1-y2-y4)	8	1000	0.9842	0.0853	0.0167
	8	1200	0.9885	0.0623	0.0100
	8	2000	0.9937	0.0344	0.0082
ACG 4 (y1-y2-y5)	7	1000	0.9848	0.0824	0.0048
	7	1200	0.9850	0.0815	0.0046
	7	1500	0.9852	0.0800	0.0058

Tablica 6. Analiza rezultata ANN za grupe neovisnih varijabli (ACG) za određivanje vrijednosti ESCE - continued

Independent variable name	Number of neurons in hidden layer	Iteration number	R^2 (Training)	SSE	
				Training	Testing
ACG 5 (y1-y2-y6)	6	1000	0.9762	0.1287	0.0365
	6	1500	0.9814	0.1008	0.0492
	6	2000	0.9777	0.1204	0.0591
ACG 6 (y1-y2-y7)	6	800	0.9774	0.1221	0.2285
	6	1200	0.9811	0.1023	0.1495
	6	2000	0.9811	0.1026	0.2651
ACG 7 (y1-y2-y5-y3)	7	600	0.9846	0.0836	0.0059
	7	800	0.9854	0.0789	0.0045
	7	1000	0.9857	0.0773	0.0053
ACG 8 (y1-y2-y5-y4)	7	1000	0.9883	0. 0635	0.0063
	7	1200	0.9893	0.0579	0.0059
	7	1500	0.9901	0.0536	0.0067
ACG 9 (y1-y2-y5-y6)	7	800	0.9870	0.0703	0.0071
	7	1000	0.9880	0.0650	0.0064
	7	1200	0.9887	0.0616	0.0069
ACG 10 (y1-y2-y5-y7)	7	1200	0.9846	0.0832	0.0129
	7	1500	0.9868	0.0719	0.0108
	7	1800	0.9882	0.0638	0.0119
ACG 11 (y1-y2-y5-y3-y4)	7	50	0.9655	0.1861	0.0347
	7	80	0.9722	0.1499	0.0208
	7	100	0.9748	0.1359	0.0230
ACG 12 (y1-y2-y5-y3-y6)	7	1200	0.9933	0.0366	0.0077
	7	1400	0.9937	0.0344	0.0069
	7	1800	0.9943	0.0309	0.0091
ACG 13 (y1-y2-y5-y3-y7)	7	1000	0.9885	0.0626	0.0088
	7	1200	0.9887	0.0612	0.0085
	7	1500	0.9894	0.0576	0.0113
ACG 14 (y1-y2-y5-y3-y6-y4)	7	100	0.9790	0.1134	0.0194
	7	200	0.9827	0.0937	0.0193
	7	400	0.9894	0.0581	0.0212
ACG 15 (y1-y2-y5-y3-y6-y7)	7	200	0.9880	0.0651	0.0089
	7	400	0.9926	0.0402	0.0076
	7	600	0.9935	0.0355	0.0088
ACG 16 (y1-y2-y5-y3-y6-y4)	7	100	0.9770	0.1245	0.0180
	7	200	0.9870	0.0708	0.0136
	7	400	0.9924	0.0412	0.0230

Analyses of independent variable groups were conducted based on ANN properties. It was seen that SCG algorithm gives better results compared to BP algorithm and, therefore, analyses were continued with SCG algorithm. Number of neurons in the hidden layer and the number of iterations varies for each independent variable group. Error values were evaluated with SSE, as shown in Equation (3).

$$SSE = \sum (y - y')^2 \tag{3}$$

SSE refers to the total square of errors, *y* refers to the real data and *y'* refers to the analysis result data.

When R^2 and the total sum of squares of errors (SSE) provided in Table 6 are examined, it can be seen that the most successful

group in the estimation of ESCE is ACG12 which has five input layers, seven hidden layers, and one output layer. This group includes the independent variables of the total indoor area (y1), wet area (y2), total exterior wall area (y3), vertical bearer area (y5), and floor area (y6).

4.3. General features of Gene Expression Programming (GEP) method and analysis of independent groups

Gene Expression Programming, (GEP), a technique developed by Ferreira, is based on genetic algorithm (GA) and genetic programming [32]. The difference between these three algorithm methods arises from the structure of chromosomes. Chromosomes are found as linear series with a fixed length in the genetic algorithm, whereas in genetic programming they are found in non-linear form and they vary in size and shape. However, chromosomes are described as linear series with a fixed length in gene expression programming, and then they are shown as non-linear simple diagrams or expression trees with various sizes and shapes [32]. Chromosomes presented as an expression tree are described in different forms and sizes by the processors (operators) found in GEP. Genetic operators such as renewal, mutation, transposition, and reintegration are used on linear chromosomes. As a result of these operators, non-linear variables in fixed numbers and lengths are converted into linear series with different sizes and forms, and functions are generated [32-34]. In this method, mathematical codes are used as the language for the genes and expression trees [32, 34-36]. This method defines all problems, from the simplest problem to the most complicated one, with an expression tree. An expression tree is comprised of mathematical statements, constants, variables, and functions [32-34,37,38]. Furthermore, expression trees can be converted in nearly all programming languages [39].

The building blocks of gene expression coding are chromosomes and expression trees. The codes of solution models in gene expression programming are made up of the genes with heads, tails and constants, and the chromosomes that contain the structure with a binding function, which binds these genes. While a solution architecture is being prepared, number of genes and head length and binding function that determine the largest size of each statement in the model, are selected [40]. Genetic operators, on the other hand, are operations carried out for the production of new generations with better qualifications using the existing population and, in this way, the scope of the search algorithm is expanded. Essentially, there are two general operators known are transposition and mutation [41]. Various genetic strategies were created in GEP through various uses of these operations and random number conversions (RNC) [42].

GeneXpro 5.0 used in this study includes 5 different training strategies listed as: optimal evolution, constant fine-tuning, model fine-tuning, subset selection, and custom [42]. Solution trees with long chromosomal structures are required for the solution of complicated problems. They are coded into smaller (such as the genes in chromosomes) structures with GEP sub-expression trees and a hierarchical structure is created [33-37, 43]. Maximum weights and depths of the sub-expression trees are calculated for each gene based on Equation (4) and Equation (5).

$$w = (n-1) \times h + 1 \tag{4}$$

Here, w refers to the weight of the sub-expression tree, h refers to the head length, while n refers to the largest value of the parameters obtained by the functions in the function set [40].

$$d = \left(\frac{h+1}{m}\right) \times \left(\frac{m+1}{2}\right) \tag{5}$$

Here d refers to the depth of the sub-expression tree, h refers to the head length, and m refers to the smallest value of the parameters obtained by the functions in the function set [40]. Binding functions consisting of addition, subtraction, multiplication and division operations are used for binding sub-expression trees together [33, 35-37, 43].

Fitness functions used in GEP demonstrate the capability of the solution genes as in the case of genetic algorithm. Fitness functions such as the mean absolute error (MAE), mean square error (MSE), root mean error (RMSE), relative square error (RSE), root relative square error (RRSE), relative absolute error (RAE), wrong balance, cost/revenue matrix and positive correlation, etc., are used in regression analyses conducted with GEP [40]. Furthermore, GEP is a genotype / phenotype genetic algorithm used as a new method to produce formulas [44].

In this study, the GEP model that was designed used the foregoing 5 variables is used as input data and updated ESCE as output data. ESCE were revised with the iteration method using 2016 as a basis. Moreover, training-testing data were determined randomly as 80 % - 20 % in the analyses.

Root relative square error (RRSE) was used as the regression analysis and fitness function within the gene expression programming. RRSE fitness function (Ei) is shown mathematically in Equation (6) and Equation (7).

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_J \tag{6}$$

Here Tj refers to the target value for j and n refers to the number of samples [42].

$$E_i = \sqrt{\frac{\sum_{j=1}^{n}\left(P_{jJ} - T_J\right)^2}{\sum_{j=1}^{n}\left(T_J - \bar{T}\right)^2}} \tag{7}$$

Table 7. Operator values according to genetic strategies*

Genetic Operators*	Optimal evolution	Constant fine-tuning	Model fine-tuning
Mutation	0.00138	0	0
Fixed-root mutation	0.00068	0	0
Function insertion	0.00206	0	0
Leaf mutation	0.00546	0	0.0140
Biased leaf mutation	0.00546	0	0.0140
Conservative mutation	0.00364	0	0.0140
Conservative fixed-root mutation	0.00182	0	0.0140
Conservative function mutation	0.00546	0	0.0140
Permutation	0.00546	0	0
Conservative permutation	0.00546	0	0.0140
Bias mutation	0.00546	0	0
Is transposition**	0.00546	0	0
Ris transposition**	0.00546	0	0
Inversion	0.00546	0	0
Tail inversion	0.00546	0	0.0140
Tail mutation	0.00546	0	0.0140
Stumbling mutation	0.00141	0	0
Uniform recombination	0.00755	0	0
Uniform gene recombination	0.00755	0	0
One-point recombination	0.00277	0	0
Two-point recombination	0.00277	0	0
Gene recombination	0.00277	0	0.01290
Gene transposition	0.00277	0	0
Random chromosomes	0.00260	0	0
Random cloning	0.00102	0.01320	0.00132
Best cloning	0.00260	0.07160	0.07160

(** The Ris and Its transposable elements of GEP are pieces of the genome that can be activated and that can be leaped to another location in the chromosome.)***
(* Definitions and the more detailed information about other parameters in the table can be obtained in [40. 42])

where P_{ij} refers to the values estimated by the model, $\overline{T}$ refers to the mean value calculated with Equation (6), T_j refers to the target value for j, and n refers to the number of samples [42].

GeneXpro includes several fixed genetic strategies. As a genetic strategy, optimal evolution, constant fine-tuning, and model fine-tuning strategies, were used in the study. The three genetic strategies used and the proportion of genetic operators in these strategies are given in Table 7.

The analyses were initially started with optimal evolution. When the best fitness value of the analysis is fixed, these three strategies were used as alternates. There are two general operators in the genetic algorithm: transition and mutation [41]. Other operators are named according to their location and methods in applying these two general operators to chromosomes or genes, cf. Table 7. The ratio of the constants for these three strategies is provided in Table 8. Moreover, the maximum fitness value of 1000 was adopted.

Table 8. Random constants*

Genetic operators*	Optimal evolution	Constant fine-tuning	Model fine-tuning
RNC mutation	0.00206	0.0328	0.0328
Constant fine-tuning	0.00206	0.0728	0.0728
Dc mutation	0.00206	0.0140	0.0140
Dc inversion	0.00546	0.0140	0.0140

(* Definitions and the more detailed information about the random constants and genetic operators in the table can be accessed in [40, 42]),

Table 9. GEP analysis adjustments

Number of chromosomes	80
Head size	20
Number of genes	1
Linking function	Multiplication
Constants per gene	10

For each run, the length of the head with the number of genes is selected. The type of the binding function, the number of genes and the length of each gene is a priority that has to be chosen for each problem. Therefore, by gradually increasing the length of the head, it is always possible to use a single chromosome. If it grows too much, the number of genes can be increased and a function can be selected for binding. In other cases, however, another binding function may be more appropriate [32]. Moreover, the number of chromosomes, head size, number of genes, and binding functions that constitute the basis of GEP architecture were determined with several trials in order to obtain the best analysis result in the study. GEP analysis adjustments are presented in Table 9.

4.4. Creating the ESCE formula

In the analysis conducted to estimate ESCE with GEP, 5 variables that have an impact on ESCE were identified as input data and updated ESCE were identified as output data. Results of the model that determines ESCE best are presented in Table 10. As a result of the model, an ESCE formula with five variables was created. This formula is shown in Equation (8).

According to Table 10, determination coefficient (R^2) for formula of ESCE was calculated as 0.90 for the training set and as 0.96 for the testing set. The Average Absolute Percent Error (MAPE) is a commonly used general measure to assess prediction accuracy. According to Table 10, the MAPE for the ESCE formula was calculated as 0.24 for the training set, and as 0.18 for the testing set. According to Lewis, the MAPE value between 0.20 and 0.50 is a reasonable forecasting value [47]. According to relevant literature, there is no exact standard for the percentage estimate to be followed for building cost estimation. However according to some studies, the accuracy of the estimation depends on the project information; It can range from + 40 % to -20 % before preliminary design, and from + 25 % to -10 % afterwards [48, 49]. A comparison of the real values obtained from the formula result is shown in figures 3, 4, and 5. Additionally, percentage errors distribution of training and testing data from the formula result is shown in figures 6 and 7.

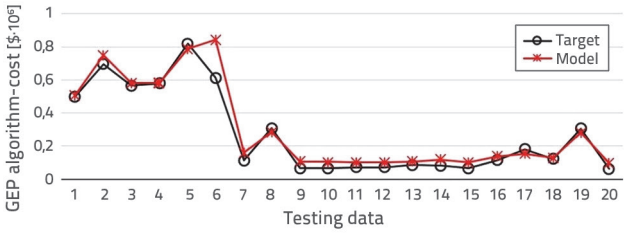
Table 10. GEP results for ESCE

	Training	Testing
R^2	0,90	0,96
Best fitness	839,05	868,98
Mean absolute error	475000	207000
Mean absolute percentage errors	0,24	0,18
Root relative square error	0,14	0,12

$$ESCE = \frac{CCI_x}{CCI_i}\left[130.47 \times \left(\frac{y_2 \times y_3 \times y_5}{y_6} + y_1\right) + 75215.80\right] \quad (8)$$

Here *ESCE* refers to the *Early-Stage Cost of Buildings*, CCI_x refers to the Construction Cost Index for the desired year, CCI_i refers to the Construction Cost Index for 2016, y_1 refers to the total indoor area, y_2 refers to the wet area, y_3 refers to the total exterior wall area, y_5 refers to the vertical bearer area and y_6 refers to the floor area.

Figure 3. GEP result of testing data for ESCE



Models designed independently and by specialized applications are needed in the construction planning and management [45]. For this purpose, the formula of the algorithm determined by using GenexPro 5.0 that estimates ESCE was created with five

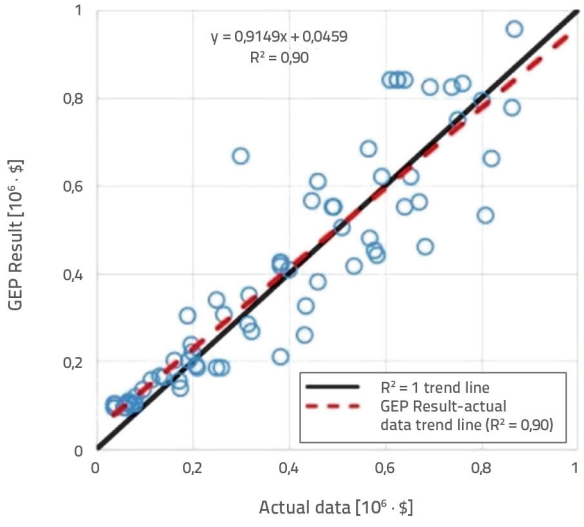


Figure 4. Training data distribution diagram for ESCE

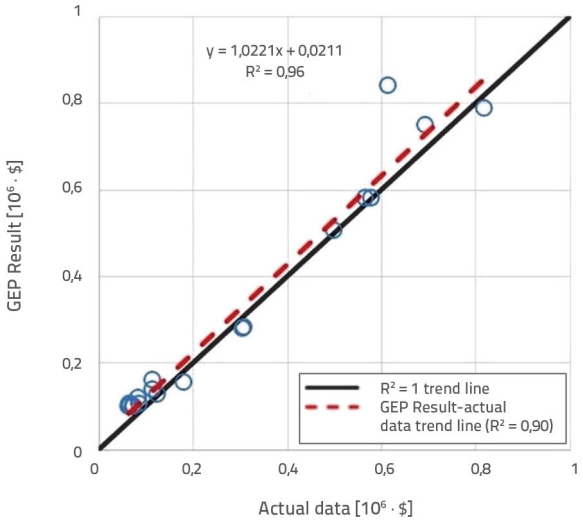


Figure 5. Testing data distribution diagram for ESCE

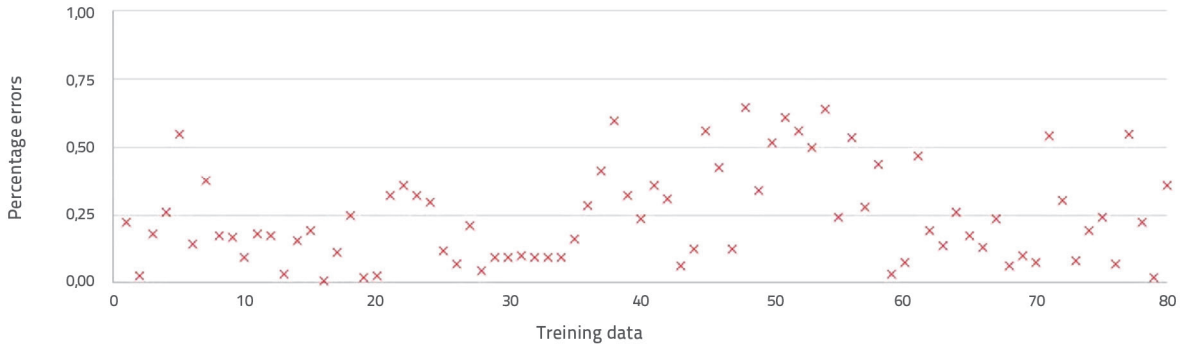


Figure 6. Percentage errors distribution of training data for ESCE

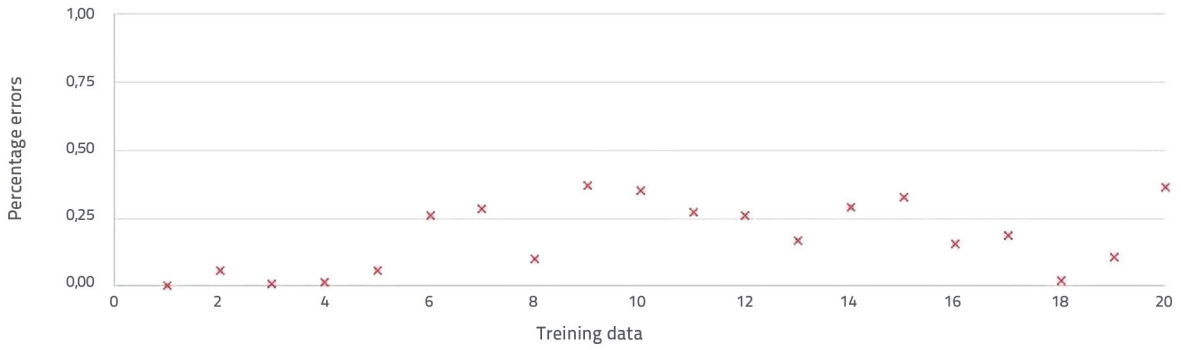


Figure 7. Percentage errors distribution of testing data for ESCE

independent variables.

The information about five independent variables that are included in static and architectural programs of the project, which differ from each other, is determined with the help of the user. As a result, the project's ESCE is estimated by the formula. The ESCE information calculated by classical method is presented in the Electronic Public Procurement Platform (EKAP) [46].

5. Case study

To see the performance of the proposed formula, a simple application was made on a real 1-storey public reinforced concrete building occupying a total floor area of 4440 square meters. This building was selected randomly among the 2019 construction tenders in EKAP. General building and tender information are given in Table 11.

Table 11. General information of a real public reinforced concrete construction

EKAP number	442.417
Bidding date	2 October 2019
Construction site	Afyonkarahisar / Turkey
ESCE in the EKAP	271064.82 $

The quantity survey study was conducted for the public reinforced concrete building using 5 independent variables. The corresponding results are given in Table 12. According to the Central Bank of the Republic of Turkey 2019 exchange rate data, 1 TL was assumed to be $ 5.7. The total indoor area of each floor was calculated by deducting the door and window lengths from the plan of each floor and multiplied by the thickness wall. The total indoor area of the building was determined by summing these values for each floor. The wet area of the building was calculated by the sum of the wet areas of each floor. The total indoor area of the building was calculated by multiplying the number of floors and floor areas. The vertical bearer area referred to the cross-sectional areas of columns and curtain elements on a single floor.

Table 12. Quantity survey result of a real public reinforced concrete construction

Total indoor area (y1)	606 m^2
Wet area (y2)	40.67 m^2
Total exterior wall area (y3)	7.84 m^2
Vertical bearer area (y5)	15.12 m^2
Floor area (y6)	606 m^2

Building Cost at an Early Stage of the real public reinforced concrete construction:

$$ESCE = 190.23 / 111.92 [130.47 x ((40.67 x 7.84 x 15.12) / / 606 + 606) + 75215.80] = 263994.40 \quad (9)$$

Here *ESCE* refers to the building cost at an early stage of the real public reinforced concrete construction, 190.23 refers to the construction cost index for September 2019, and 111.92 refers to the construction cost average index for 2016.

The ESCE of the public reinforced concrete building calculated by the Employer (public institutions) amounted to 271064.82 $, while the corresponding amount calculated using the proposed formula was 263994.40 $. When these two results are examined, it can be seen that the accuracy of the ESCE estimation using the proposed formula amounts to 97 %.

6. Discussion and conclusion

The Early-Stage Cost of Buildings (ESCE) is one of the most critical components of an employer's and contractor's budget accounts on a construction project. Employers wish to obtain accurate results in cost estimations in the shortest time to arrange the budgets of their investments, whereas the contractors wish to calculate their profits accurately. The traditional estimation of building cost based on quantity survey calculation for the entire project is quite accurate but is also time-consuming. At the preliminary design stage or where the bidding period is limited, building costs may need to be calculated in an expedited manner. According to Kim et al. (2004), it would be appropriate to create a genetic algorithm-based model to obtain optimum cost estimation parameters with optimum NN architecture in the building cost estimation [50].

In this study, a formula was proposed to estimate the Early-Stage Cost Estimation of Building Construction Projects (ESCE) in a rapid, easy and accurate manner by using the data selected according to the construction design and by employing the ANN and GEP methods. A quantity survey study was conducted on one hundred construction projects tendered between 2011 and 2016 in relation to independent variables determined to have an impact on building costs, and a data set was created. The total floor area of the projects used in the data set varied between 141 m^2 to 7947 m^2. According to Khamis et al. (2005), outliers in the training and testing data reduce the accuracy of the model [51]. For this reason, determining the limit values for the total floor area of the projects prevented formation of extreme values and increased the learning performance of the model. The data set was examined with ANN analyses in order to determine the variables affecting the ESCE. As a result of these analyses, the total indoor area (y1), wet area (y2), total exterior wall area (y3), vertical bearer area (y5), and floor area (y6) variable groups were used to estimate ESCE, and successful results were achieved. The ESCE estimation coefficient of this group was found to be 0,99 for the training set. At the next stage, a model configuration was made with GEP where the independent variables found as a result of the ANN analysis were input data for GenexPro 5.0. The formula of the model (formula of ESCE-Eq. (8)) was created with these five independent variables. The determination coefficient (R^2) for this ESCE formula was calculated as 0.90 for the training set and as 0.96 for the testing set. The testing set MAPE value for the ESCE formula was calculated as 0.18 which is within reasonable forecasting value limits.

In addition, the correlation and linear regression analysis was performed on the data set in order to compare this model (this formula) with regression analysis, which is a classical method. As a result of the analysis, it was determined that the building importance coefficient, vertical bearer area, and building height parameters, were effective in the early stage building cost estimation and R^2 value was found to be 0.77. Kim et al.[7], Cho et al. [16], Latief et al. [17], Kim et al [50], showed that artificial intelligence techniques (such as ANN, Neuro Fuzzy) performs better than regression analysis, which is a classical method for estimating construction costs. Finally, verifications were made on a case study to see the efficiency of the formula. The following conclusions were drawn in this research:

- The proposed formula alleviates the burden of long-lasting quantity surveys for reinforced concrete construction projects.
- The proposed formula estimates ESCE rapidly and easily.
- It was observed during the study that the quantity survey calculation of these projects even in ESCE varied from one case to another. The use of the proposed formula in the early-stage building cost calculations is important not only for faster and easier cost calculation but also to prevent any differences that may arise due to the individual making the calculations.
- On the other hand, the results of the research showed the validity of using ANN and GEP together in the calculation of ESCE.
- It was proven that the ESCE formula with satisfactory precision can be created.
- Moreover, a significant contribution is made to the literature, in the sphere of facilitating an easy, rapid and accurate estimation of ESCE on construction projects.

Acknowledgement

The study benefited from the support by Sakarya University-Scientific Research Projects Unit (Project No: 2016-07-12-004)

REFERENCES

[1] Park, H.K.: Cash flow forecasting in construction project, KSCE J. Civ. Eng., 8 (2004), pp. 265–271.

[2] Budak, O.: Solution proposals for the application problems of public procurement law at numbered 4734, M.S. Thesis, Istanbul Technical University, (2006).

[3] Boussabaine, A.H., Elhag, T.M.S.: An artificial neural system for cost estimation of construction projects, Proc. 14th Annu. ARCOM Conf., Reading, United Kingdom, (1998).

[4] Lowe, D.J., Emsley, M.W.: A harding, predicting construction cost using multiple regression techniques, J. Constr. Engıneerıng Manag., 132 (2016), pp. 750–758.

[5] Hwang, S.: Dynamic regression models for prediction of construction costs, J. Constr. Eng. Manag., 135 (2009), pp. 360–367.

[6] Cheng, M.Y., Tsai, H.C., Sudjono, E.: Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry, Expert Syst. Appl., 37 (2010) 6, pp. 4224-4231, doi:10.3923/jai.2011.63.7510.1016/j.eswa.2009.11.080.

[7] Kim, G., Shin, J., Kim, S., Shin, Y.: Comparison of school building construction costs estimation methods using regression analysis neural network, and support vector machine, J. Build. Constr. Plan. Res., 1 (2013), pp.1-7, doi:10.4236/jbcpr.2013.11001.

[8] Bostancioglu, E.: The factors that affecting residential buildings cost in pre-design phase and a cost estimating method that is based on these factors, Ph.D. Thesis. Istanbul Technical University. (1999)

[9] Gunaydın, M.H., Dogan, Z.S.: A neural network approach for early cost estimation of structural systems of buildings, Int. J. Proj. Manag., 22 (2004) 7, pp. 595-602, https://doi.org/10.1 016/j.ijproman.2004.04.002.

[10] Akinbingol, M., Gultekin, A.T.: A cost management model proposal for cost planning and supervision in building production construction, Gazi University Journal of Engineering and Architecture Faculty, 20 (2005), pp. 499–505.

[11] Dogan, S.Z., Arditi, D., Gunaydın, H.M.: Determining attribute weights in a cbr model for early cost prediction of structural systems, J. Constr. Eng. Manag., 132 (2006), pp. 1092-1098, https://doi.org/10.1061/(ASCE)0733-9364(2006)132:10(1092).

[12] Nan, J., Choi, J.W., Choi, H., Kim, J.H.: A study on estimating construction cost of apartment housing projects using genetic algorithm support vector regression, KJCEM. 15 (2014), pp. 68-76.

[13] Sonmez, R.: Range estimation of construction costs using neural networks with bootstrap prediction intervals, Expert Syst. Appl., 38 (2011) 8, pp. 9913-9917, https://doi.org/10.1016/j.eswa.2011.02.042.

[14] Arafa, M., Alqedra, M.A.: Early stage cost estimation of building construction projects using artifical neural networks, J. Artifical Intell., 4 (2011) 1, pp. 63-75, doi:10.3923/jai.2011.63.75.

[15] Kuruoglu, M., Yonez, E., Topkaya, E., Celik, L.Y.: Comparison of preliminary cost estimation methods used in the construction sector, e-Journal New World Sci. Acad., 7 (2012), pp. 263–272.

[16] Cho, H.G., Kim, K.G., Kim, J.Y., Kim, G.H.: A comparison of construction cost estimation using multiple regression analysis and neural network in elementary school project in the early stages of a construction project, Journal of the Korea Institute of Building Construction, 13 (2013) 1, pp. 66-74, doi:10.5345/JKIBC.2013.13.1.066.

[17] Latief, Y., Wibowo, A., Isvara, W.: Preliminary cost estimation using regression analysis incorporated with adaptive neuro fuzzy inference system, Int. J. Technol., 1 (2013), pp. 63–72.

[18] El-sawalhi, N.I., Shehatto, O.: A neural network model for building construction projects cost estimating, Journal of Construction Engineering and Project Management, 4 (2014), pp. 9-16.

[19] Bayram, S., Ocal, M.E., Oral, E.L., Atis, C.D.: Comparison of unit price method and unit area cost method for construction cost estimation, Journal of Polytechnic, 19 (2016), pp. 175–183.

[20] Coloma, J.F., Valverde, L.R., García, M.: Estimation of construction costs of rustic homes through artificial neural networks, Informes de la Construcció, 71 (2019), pp. 554.

[21] Dimitrijević, B., Stojadinović, Z., Marinković, D., Dimitrijević, M.: Influence of structural system on the construction time and cost of residential projects, GRAĐEVINAR, 71 (2019) 8, https://doi.org/10.14256/JCE.2315.2018

[22] Naser, M.Z., Abu-Lebdeh, G., Hawileh, R.: Analysis of RC T-beams strengthened with CFRP plates under fire loading using ANN, Construction and Building Materials (2012), doi:10.1016/j.conbuildmat.2012.07.001

[23] Leśniak, A., Juszczyk, M.: Prediction of site overhead costs with the use of artificial neural network based model, Archives of Civil and Mechanical Engineering (2018), doi.org/10.1016/j.acme.2018.01.014

[24] TUIK, Construction Cost Index and Rate, 07 December 2019.

[25] Naser, M.Z.: Deriving temperature-dependent material models for structural steel through artificial intelligence, Construction and Building Materials (2018), doi:10.1016/j.conbuildmat.2018.09.186.

[26] Caglar, N.: Artifical neural networks in dynamic analysis of buildings, Ph.D. Thesis. Sakarya University. (2001).

[27] Oztemel, E.: Artificial neural networks, Papatya Publisher, Istanbul, (2012).

[28] Sagiroglu, S., Besdok, E., Erler, M.: Engineering artificial intelligence applications 1/ artificial neural networks, Ufuk Publisher, Istanbul, (2003).

[29] Hamzacebi, C.: Artificial neural networks, Ekin Publisher, Bursa, (2011).

[30] Eren, B., Turp, S.M.: Estimation of Nickel (Ii) Ion Removal Efficiency From Seepage Water With Artificial Neural Networks, e-Journal of New World Sciences Academy, 6 (2011), pp. 398–405.

[31] Caglar, N.: Neural network based approach for determining the shear strength of circular reinforced concrete columns, Construction and Building Materials, (2009), doi:10.1016/j.conbuildmat.2009.06.002.

[32] Ferreira, C.: Gene expression programming: a new adaptive algorithm for solving problems, Complex Syst., 13 (2001), pp. 87–129.

[33] Ferreira, C.: Function finding and the creation of numerical constants in gene expression programming, 7th Online World Conf. Soft Comput. Ind. Appl. Bristol, (2002).

[34] Saridemir, M., Kara, I.F.: Estimation of the tensile strength of the fiber-reinforced concrete containing silica fume by GEP, Journal of Nigde University Engineering Science 5 (2016), pp. 208-217.

[35] Kara, I.F.: Prediction of shear strength of frp-reinforced concrete beams without stirrups based on genetic programming, Adv. Eng. Softw., 42 (2011), pp. 295–304.

[36] Nazari, A., Riahi, S.: Prediction split tensile strength and water permeability of high strength concrete containing TiO2 nanoparticles by artificial neural network and genetic programming, Compos. Part B., 42 (2011), pp. 473–488.

[37] Ferreira, C.: Genetic representation and genetic neutrality in gene expression programming, Adv. Complex Syst. 5 (2002), pp. 389–408.

[38] Severcan, M.: Prediction of splitting tensile strength from the compressive strength of concrete using GEP, Neural Comput. Appl. 21 (2012), pp. 1937–1945.

[39] Sette, S., Boullart, L.: Genetic programming: principles and applications, Eng. Appl. Artif. Intell., 14 (2001), pp. 727–736.

[40] Ferreria, C.: From GeneXproTools Documentation: a gepsoft web resource, (2017). www.gepsoft.com/GeneXproTools/Regression.ht

[41] Isci, O., Korukoglu, S.: An Application in genetic algorithm approach and steering research, Management and Economics, 10 (2003).

[42] Gepsoft, GeneXproTools., www.gepsoft.com, 10 May 2018.

[43] Sarıdemir, M.: Genetic programming approach for prediction of compressive strength of concretes containing rice husk ash, Constr. Build. Mater., 24 (2010), pp. 1911–1919.

[44] Hadianfard, M.A., Jafari, S.: Prediction of lightweight aggregate concrete compressive strength using ultrasonic pulse velocity test through gene expression programming, Scientia Iranica A, 23, (2016), pp. 2506-2513.

[45] Scherer, R.J., Schapke, S.E.: A distributed multi-model-based management information system for simulation and decision-making on construction projects, Advanced Engineering Informatics (2011), doi:10.1016/j.aei.2011.08.007.

[46] EKAP. Republic of Turkey Electronic Public Procurement Platform Official Website, 07 December 2019.

[47] Lewis, C.D.: Industrial and business forecasting methods. London: Butterworths, 1982.

[48] Oberlender, G.D.: Project Management for Engineering and Construction, McGraw-Hill, Inc., 1993.

[49] Enshassi, A., Mohamed, S., Madi, I.: Cost estimation practice in the Gaza Strip: A case study, The Islamic University Journal, 15 (2007), pp. 153-176.

[50] Kim, G.-H., An, S.-H., Kang, K.I.: Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning, Building and Environment, 39 (10) (2004), pp. 1235-1242.

[51] Khamis, A., Ismail, Z., Haron, K., Mohammed, A.T.: The effects of outliers data on neural network performance, Journal of Applied Science, 5 (2005) 8, pp. 1394-1398.