Contents lists available at ScienceDirect

Heliyon

journal homepage: www.heliyon.com

Heliyon

Intelligent methodology for project conceptual cost prediction

Haytham H. Elmousalami *

Department of Construction and Utilities, Faculty of Engineering, Zagazig University, Egypt

ARTICLE INFO

Keyword:
Computer science

ABSTRACT

Developing a reliable parametric cost model at the conceptual stage of the project is crucial for projects managers and decision makers. Several methodologies exist to develop a conceptual cost model. However, many gaps exist in the current methodologies such as depending only on experts ‘opinions and questionnaire survey to identify the project features, key cost drivers and developing deterministic predictive models without taking uncertainty nature into consideration. The main contribution of this study is developing an intelligent methodology for predicting the project cost at the conceptual stage. The proposed methodology can automatically identify key cost drivers and maintain uncertainty to predicted cost. Field canals improvement projects (FCIPs) are used as a case study to validate the proposed methodology. The selected methodology has applied quantitative approaches to identify the key cost drivers. In addition, the methodology has applied a genetic fuzzy model that automatically generates fuzzy rules to automatically predict the conceptual cost. Moreover, the results show a superior performance of the genetic fuzzy model than the traditional fuzzy model. In addition, this study presents a publicly open dataset for FCIPs to be used for future models validation and analysis.

1. Introduction

Conceptual cost estimate occurs at 0%–2% of the project completion (PMI, 2008; AACE International recommended practices, 2004). At the conceptual stage of the project, cost prediction is a critical process where crucial decisions about the project depend on it and limited information about the project is available (Hegazy and Ayed, 1998). Feasibility study cost estimate occurs at 1%–15% of the project completion based on the parametric model where its accuracy varies from -30% to +50% (AACE International recommended practices, 2004). Capacity factored model, analog model (near neighbor) and parametric model are conducted to perform such conceptual estimate where its accuracy varies from -50% to +100% (AACE International recommended practices, 2004). However, parametric cost estimate resents more accurate results than capacity factored model and analog model where parametric cost estimate deeply construct cost estimating relationships (CERs) between cost and cost predictors (PMI, 2008; AACE International recommended practices, 2004).

Parametric cost modeling is creating a model based on key cost drivers extracted from experts’ experience or the collected past cases by conducting statistical analyses(Dell’Isola, 2002). The first and the most important step in the parametric cost model development is key cost drivers' identification. Many previous studies have conducted qualitative approaches such a questionnaire survey [(Marzouk and Ahmed, 2011) (ElSawy, et al, 2011)], Delphi method [(Liu, 2013), (Hsu et al., 2010)] or Analytical hierarchy method [(Liu, 2013), (Manoliadis et al., 2009)] to identify key cost drives. Based on Fifty-two factors of Building construction in Egypt, ten cost drivers have been selected by a questionnaire survey of experts for artificial neural networks (ANNs) cost model (ElSawy, et al, 2011). A questionnaire survey has been operated to identify the input variables for ANNs for cost deviation where 36 factors have been identified (Attalla and Hegazy, 2003). A questionnaire survey has been conducted to determine significant parameters for ANNs cost prediction model for tunnel construction in Greece (Petroutsatou et al., 2012). However, the main gap of these studies is depending only on experts' opinions, interviews, and questionnaire survey to identify key cost drivers where experts' opinions may produce biased or wrong opinions. In addition, interviewing experts is a time and effort consuming process.

The second step in the parametric cost model development is formulating mathematical relationships among key cost drivers and the cost of the project within the acceptable prediction accuracy. Many previous studies have applied artificial intelligence (AI) techniques and machine learning (ML) models such as ANNs, regression model, case-based reasoning (CBR), hybrid models such as neural-fuzzy models, and evolutionary computing (EC) such as genetic algorithm (GA) and

* Corresponding author.
E-mail address: Haythamelmousalami2014@gmail.com.

https://doi.org/10.1016/j.heliyon.2019.e01625
Received 18 November 2018; Received in revised form 30 March 2019; Accepted 29 April 2019

genetic fuzzy models.

Building information modeling (BIM) can feed data for cost estimation where a predictive ML model such as regression model or ANNs can predicted the project's cost on a macro level (Juszczyk, 2017). ANN has been applied for cost estimation of sports fields where the general applicability of ANNs model has been investigated (Juszczyk et al., 2018). ANNs model has been conducted to early cost estimate of building projects for reinforced concrete buildings with acceptable performance (Ambrule and Bhirud, 2017). A semilog regression model has performed to develop cost models for residential building projects in German with a prediction accuracy of 7.55% (Stoy et al., 2012). Based on 92 building projects, ANNs and supportive vector machine (SVM) have been used to predicted cost and schedule success at the conceptual stage. Such a model has a prediction accuracy of 92% and 80% for cost success and schedule success, respectively (Wang et al., 2012).

Based on 657 building projects in Germany, a multistep ahead approach is conducted to increase the accuracy of the model's prediction (Dursun and Stoy, 2016). Analytic hierarchy process (AHP) has incorporated into CBR to build a reliable cost estimation model for highway projects in South Korea (Kim et al., 2013). CBR has been proposed for estimation the preliminary costs of sports field construction based on 16 predictors using 143 construction projects. Different calculations were conducted to formulate the case similarity based on quantitative and qualitative data where the final total error was 14% at the early stage (Leśniak and Zima, 2018). Prediction performance of a cost prediction model has been improved by 17.23% and 4.39% for business facilities model and multi-family housings model, respectively. CBR technique applied multiple regression analysis (MRA) technique in the revision phase of the CBR technique where this integration can reliably predict the conceptual cost estimate (Jin et al., 2012).

Parametric cost modeling is to develop a model based on logical or statistical relations of the key cost drivers extracted by conducting qualitative techniques (Elmousalami et al., 2018a) or statistical analyses such as factor analysis (Marzouk, and Elkadi, 2016) or stepwise regression technique (Elmousalami et al., 2018b). Based on more than 1,400 projects, a multilayer of ensemble methods has been developed for forecasting the unit price bids of resurfacing highway projects (Cao et al., 2018). Wang and Ashuri (2016) have applied a random tree model for construction cost index prediction. Williams and Gong (2014) have built a stacking ensemble learning and text mining to estimate the cost overrun using the project contract document where the accuracy was 44%. Building Information Modelling (BIM) can automate cost estimation process and improve inaccuracies where New Rules of Measurement (NRM) for cost estimation can be extracted for automatic cost estimate based on a 4D BIM modeling software (Kim et al., 2019).

Arabzadeh et al. (2018) have developed ANNs, regression and hybrid models for cost estimation of spherical storage tanks. The results indicated that ANN was more accurate than hybrid regression model and hybrid ANNs was more accurate then single ANNs. Linear and multiple regression models have been counted to predict the preliminary estimate of road projects in Nigeria at the early stage (Ogungbile et al., 2018). However, the whole

The collected data set was only 50 for seven predictors where it is not sufficient data size to train regression models. Zhang et al. (2018) have converted time series model into a graph to forecast the construction cost index where the application showed its ability to provide more accurate estimations.

However, the main gap of these studies is developing deterministic predictive models without taking uncertainty nature into account where adding uncertainty nature to the predicted values improves the quality and reliability of the developed models (Zadeh, 1965, 1973). Therefore, the purpose of the study is to formulate a general methodology that can be conducted to develop a reliable parametric cost prediction model at the conceptual stage of the project. The proposed methodology should handle the following objectives:

1 The first research objective is to automatically determine the cost drivers of FCIPs based only on the collected data based on quantitative approaches without depending on experts' opinions such as questionnaire survey, Delphi methods, or AHP (Elmousalami et al., 2018a).
2 The second research objective is to conduct fuzzy logic theory to prediction modeling to maintain uncertainty to the predicted values.
3 The third objective is to use the hybrid fuzzy modeling to overcome the fuzzy rules generation problem.

2. Research methodology

As illustrated in Fig. 1, the first step in the proposed methodology is a literature review of the previous practices for key drivers' identification. ML techniques cannot work without data. Accordingly, the second step is data collection of FCIPs historical cases. The third step is model development based on the quantitative approach for key cost drivers identification and fuzzy modeling for cost prediction. The quantitative approach is applying exploratory factor analysis (EFA), regression methods, correlation matrixes, and hybrid methods. The fuzzy modeling can be traditional or hybrid modeling. The final step in model development is model validation.

3. Quantitative approaches for key cost drivers' identification

The purpose of variables selection is to improve the prediction accuracy and provide a better understanding of collected data (Guyon and Elisseeff, 2003). The poor variables selection can decrease model precision. Moreover, identifying cost and project features are time-consuming, difficult, and requires expert knowledge.

Applied ML is basically feature engineering where several methodologies can be applied to select the key predictors. This could be achieved by deleting both irrelevant predictors and redundant predictors based on statistical criteria such as R-squared or P–values (Ratner, 2010).

After collecting relevant data which represents all variables, statistical methods can be used to analyze data and screen such variables. This study aims to conduct ML models that can accurately and automatically identify the key cost drivers with the fewer number of the project parameters such as construction year or project's area. In addition, the study has conducted and compared different statistical techniques such as exploratory factor analysis, regression methods, and correlation matrix scanning to identify the key cost drivers. Accordingly, the current study will conduct these techniques and compare them.

3.1. Applying factor analysis

Marzouk and Elkadi (2016) identified cost drivers that influence construction costs of water treatment plants. Cost drivers have been determined through Descriptive Statistics Ranking (DSR) and EFA. Principal component analysis (PCA) with varimax rotation through five iterations were used to minimize multicollinearity problem. A total of *33* variables were reduced to eight components while using Cattell's Scree test reduced variables to four components. Woldesenbet et al., (2012) have conducted the factor analysis of a covariance and correlation matrix to investigate the significance and correlation of critical factors affecting preliminary cost of roadway projects. Alroomi et al. (2012) identified 23 core estimating competencies classified into skills, knowledge, and personal attributes and also quantified the degree where new estimators lack each competency. The factor analysis has grouped these *23* competencies into seven different factors by using the factor analysis method.

3.2. Applying regression methods

Stoy et al. (2012) developed conceptual cost models for German residential building project. Historical data were randomly sampled from the Building Cost Information Centre. A total of *75* residential projects

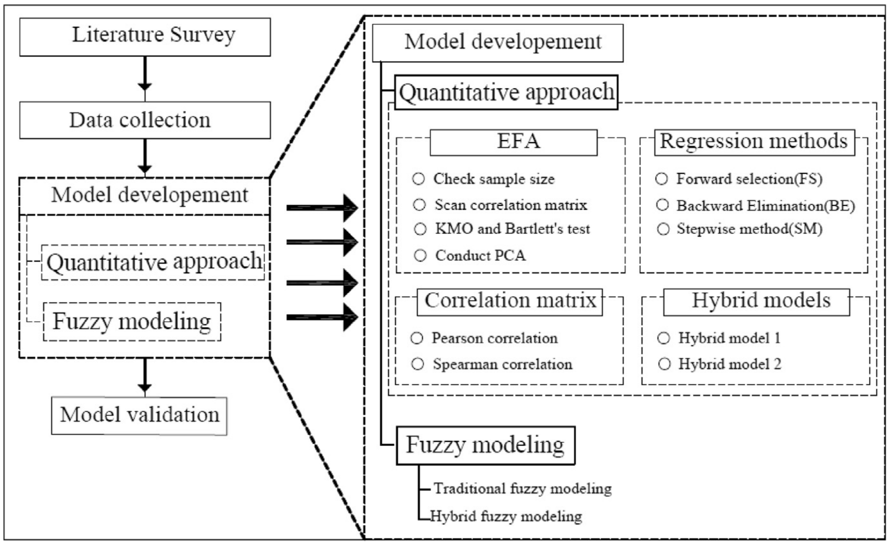


Fig. 1. Research methodology.

has been sampled. Multi-collinearity and singularity problems have been detected and eliminated where the most significant predictors were compactness, the percentage of openings and height of the building for the cost of external walls. These parameters were determined by a backward regression method. Lowe et al., (2006) described the development of linear regression models to predict the construction cost of buildings, based on *286* sets of data collected in the United Kingdom. Both forward and backward stepwise regression analyses were performed to produce a total of six models. Forty-one independent variables have been identified and classified either as project strategic, site-related or design related.

3.3. Applying correlation methods

Yang (2005) presented a general method to incorporate correlations between cost elements in the process of cost estimation. Yang (2005) proposed a simulation-based method to estimate project cost while considering correlations between cost elements where it can automatically adjust an infeasible correlation matrix into a close and feasible one very efficiently. The proposed method has first checked the feasibility of the correlation matrix, adjusts it if necessary, then has used the correlations to generate correlated multivariate random vectors to generate outcomes of the cost elements. The method was applied to a full data set of 216 British office buildings. The application result indicated that the impact of correlations was significant and may cause serious problems if neglected. Ranasinghe (2000) used the correlation matrix for selecting the input project cost variables based on *70* German residential properties. Stoy et al. (2008) used a series of independent variables for early estimation of building construction cost of residential buildings by regression analysis. These variables serve as cost drivers of a project. As illustrated by the literature survey, most studies conducted factor analysis, regression analysis, or correlation matrix to select the key cost drivers.

4. Fuzzy modeling for prediction

Once the key cost drivers have been identified, a prediction model is required to convert the selected cost drivers to the final output (conceptual cost of the project). Fuzzy logic (FL) is to model human reasoning taking uncertainties possibilities into account where incompleteness, randomness, and ignorance of data are represented in the model (Zadeh, 1965, 1973).

Linguistic variables are labels of fuzzy subsets whose values are words or sentences (Zadeh, 1976). Such linguistic terms mean approximation of system features which cannot be represented precisely by quantitative terms. For example, a project cost is a linguistic variable where consists of high cost, medium cost, and low cost. “High cost” is a linguistic term of the project cost, for example, compared with an exact numeric value ‘33 million dollars’. If–Then rule statements are utilized to formulate the conditional statements that develop FL rules base system. A single fuzzy If–Then rule can be represented by the following:

If *fuzzy proposition (x is A_1)>* ***Then*** *fuzzy proposition (y is B_2)>*

Where x is an input parameter and A_1 is a membership function (MF) of x, and y is an output parameter and B_2 is an MF of y. Rule-based systems are systems that have more than one rule to represent human logic and experience to the developed system. Aggregation of rules is the process of developing the overall consequent from the individual consequents added by each rule (Siddique and Adeli, 2013).

As shown example in Fig. 2, there are two parameters X_1 and X_2 where $\mu X_1 = \{a_1, b_1, c_1, d_1\}$, $\mu X_2 = \{a_2, b_2, c_2, d_2\}$, $\mu Y = \{a_y, b_y, c_y, d_y\}$ and the fuzzy system consists of two rules as following:

Rule 1: IF x_1 is a_1 AND x_2 is c_2 THEN y is a_y.
Rule 2: IF x_1 is b_1 AND x_2 is d_2 THEN y is b_y.

Where two inputs are used $\{X_1=4, X_2=6\}$. Such two inputs intersect with the antecedents MF of the two rules where two consequents rules are produced $\{R_1$ and $R_2\}$ based on minimum intersections (Siddique and Adeli, 2013). The consequent rules are aggregated based on maximum intersections where the final crisp value is 3. The aggregated output for R_i rules are given by.

Rule 1: $\mu R_1 = \min [\mu a1 (x1) \text{ and } \mu c2 (x_2)]$
Rule 2: $\mu R_2 = \min [\mu b1 (x1) \text{ and } \mu d2 (x_2)]$
Y: fuzzification $[\max [R_1, R_2]$

Fuzzification is converting a numeric value (or crisp value) into a fuzzy input. Conversely, defuzzification is the opposite process of fuzzification where the defuzzification is the conversion of a fuzzy quantity into a crisp value. Max-membership, the center of gravity, weighted average, mean-max, different defuzzification and center of sums are different defuzzification methods (Runker, 1997).

4.1. Traditional fuzzy cost model

Many parametric cost models have been developed based on fuzzy

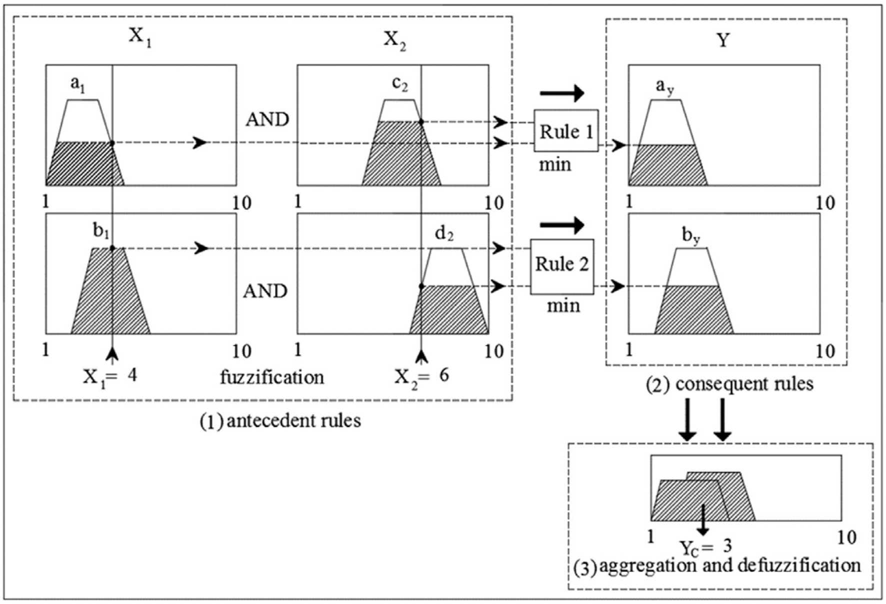


Fig. 2. Fuzzy rules firing.

theory to utilize uncertainty concepts for a cost estimate. Based on 98 examples, Ahiaga-Dagbui et al.,2013) have developed a cost model for water infrastructure projects where a combination of ANNs and fuzzy set theory are incorporated to develop a more accurate model where mean absolute percentage error (MAPE) was 0.8%. Based on 568 Towers, a four input fuzzy clustering model and sensitivity analysis are conducted for estimating telecommunication towers with acceptable MAPE (Marzouk and Alaraby, 2014). Shaheen et al. (2007) have proposed the use of fuzzy numbers for cost range estimating and claimed the fuzzy numbers for fuzzy scheduling range assessment. Shreenaath et al. (2015) have conducted a statistical fuzzy approach for prediction of construction cost overrun. Based on 60 respondents and relative important index (RII) scale, five factors are selected of 54 factors to be used as fuzzy model inputs. In addition, the model is validated by four case studies.

Marzouk and Amin (2013) have developed an ANNs model for predicting construction materials prices, whereas the FL model is applied to determine the importance degree of each material for ANNs model. Such a model has an acceptable accuracy in the training and testing phases. The FL model is developed for satellite cost estimation. Such model works as a fuzzy expert tool for cost prediction based on two input parameters (Karatas and Ince, 2016).

By surveying the past literature, the studies have developed fuzzy systems without mentioning the method of fuzzy rules generation or the fuzzy rules has been developed based on experts' experience. Determining the fuzzy rules is the main gap of the previous studies. Therefore, a new trend evolves to solve this problem such as developing hybrid fuzzy modeling for the cost estimate purposes such as evolutionary-fuzzy modeling.

4.2. Evolutionary fuzzy systems

Many approaches exist for evolutionary fuzzy hybridization [(Angelov, 2002) (Pedrycz, 1997) (Herrera et al, 1996)]. Genetic algorithm (GA) is an effective evolutionary algorithm (EA) used for search and optimization applications (Goldberg, 2002). Selecting the MF for each linguistic variable and developing fuzzy IF-Then rules are the main problems in fuzzy system development (Cordon et al., 2001). Traditionally, an expert is consulted to define such fuzzy rules or the fuzzy designer can use the trial and error approach to map the fuzzy rules and MFs. However, such an approach is time-consuming and does not guarantee the optimal set of fuzzy rules.

Moreover, the number of fuzzy IF-Then rules increase exponentially by increasing the number of inputs, linguistic variables or a number of outputs. In addition, the experts cannot easily define all required fuzzy rules and the associated MFs. In many engineering problems, EA has been conducted to automatically develop fuzzy rules and MFs to improve the system performance (Chou, 2006; Kwon and Sudhoff, 2006; Loop et al., 2010).

As illustrated in Fig. 3, there are three modules for EA fuzzy systems: A, B and C. Module (A) used EA to develop the optimal fuzzy rules, and module (B) used EA to determine the optimal MFs. whereas module (C) used EA for determining both MFs and Fuzzy rules. Karr and Gentry (1993) have applied GA for tuning and computing MFs for FL controllers to improve the system performance. Ishibuchi et al., (1999) have developed a genetic fuzzy system for classification. Linkens and Nyongesa (1995 a, b) have simplified and generated linguistic and fuzzy rules using GA. Fuzzy rules can be fixed and the MFs are tuned, on the other hand, fuzzy rules can be tuned while the MFs can be fixed. Homaifar and McCormick (1995) have developed a genetic fuzzy system which can simultaneously generate both rule sets and MFs to eliminate the human need for fuzzy system design.

The fuzzy rules base or IF–Then rules are the core of any fuzzy system that consists of a set of IF–Then rules. The performance of any fuzzy system mainly depends on the rule base which are IF–Then rules. Evolutionary learning is a suitable technique as it can incorporate prior knowledge of the developed system (Belarbi and Titel, 2000; Bonarini and Trianni, 2001). The prior knowledge may be in the form of linguistic variables, MF parameters, and fuzzy rules. EA such as GA can be incorporated into the fuzzy system to obtain the learning capability (Hinton and Nowlan, 1987; Bonarini and Trianni, 2001). Evolutionary learning can be merged in the fuzzy system to optimize its parameters such as MF parameters, fuzzy rules, and the number of rules. Structure learning (i.e., rule-based learning) and parameter learning (i.e., MF learning) are the two kinds of fuzzy system learning.

Two main approaches exist to conduct evolutionary fuzzy system: Michigan approach (Casillas et al., 2007; Bonarini, 1996; Ishibuchi et al., 1999), Pittsburgh approach (Hoffmann and Pfister, 1997). Michigan approach is to represent each chromosome as a single rule, whereas the rule set is the entire population. The objective of the EA is to select the optimal subset of chromosomes that represents the optimal set of rules. The Michigan approach is outlined as follows:

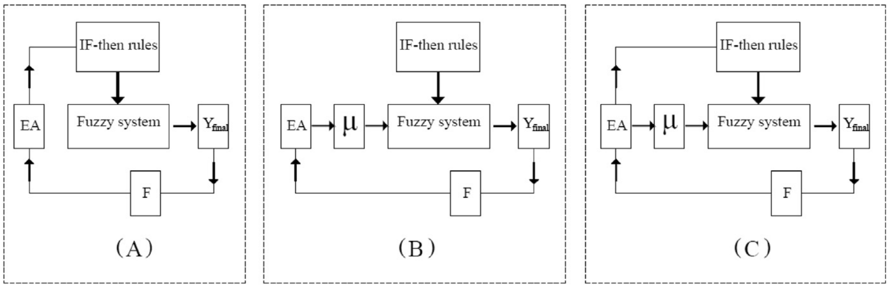


Fig. 3. Evolutionary fuzzy systems (A) generating IF-Then rules system (B) generating membership functions system (C) generating both IF-Then rules and membership functions system.

1. Generate randomly an initial population of fuzzy IF-Then rules.
2. Select a sample from the developed population and evaluate the fitness of the rules of the selected sample.
3. Generate new individuals of fuzzy IF-Then rules by genetic operators.
4. Replace individuals with new individuals of the population.
5. Continue until no further improvement of system performance.

On the other hand, Pittsburgh approach represents each chromosome as a set of fuzzy rules where the number of rules constant (Herrera, 2008). Similarly, Pittsburgh approach can be conducted such as the Michigan approach.

4.3. *Hybrid fuzzy cost model*

Zhai et al. (2012) have created an improved fuzzy system which is established based on fuzzy c-means (FCM) to solve the problem of fuzzy rules generation. This model has produced better results for scientific cost prediction. Cheng et al. (2009) have incorporated computation intelligence models such as ANNs, FL, and EA to make a hybrid model which improves the prediction accuracy. Similarly, Chen et al. (2010) have developed an evolutionary fuzzy hybrid neural network for. Conceptual cost estimates. FL is used for fuzzification and defuzzification for inputs and outputs, respectively. GA is utilized for optimizing the parameter of the model such as NN layer connections and FL membership. As a result, an evolutionary fuzzy neural model has been developed for conceptual cost estimation for building projects.

Zhu et al, (2010) have conducted an evolutionary fuzzy neural network model for cost estimation based on eighteen examples and two examples for training and testing, respectively. The previous study has an insufficient sample size for model training where Green (1991) has recommended that $50 + 8k$ may be the minimum sample size, where k is the number of predictors. GA is conducted for model optimization and to avoid sinking into local minimum results. Cheng and Roy (2010) have developed a hybrid AI system based on SVM, FL, and GA for decision making construction management. The system has applied the FL to handle uncertainty to the system, SVM to map fuzzy inputs and outputs, and GA to optimize the FL and SVM parameters. The objective of such a system is to produce accurate results with less human interventions, where MF shapes and distributions can be automatically mapped. The past studies have developed cost estimate models based on hybrid fuzzy systems. The objective of the hybrid systems is to develop reliable fuzzy models that have no limitations of the traditional fuzzy model such as fuzzy rules generation.

5. Application to FCIPs

In this section, the proposed method is applied to the cost estimation of Field canals improvement projects (FCIPs) in Egypt.

5.1. *Case background and data collection*

FCIPs are one of the main projects in Irrigation Improvement Projects (IIPs). The strategic aim of these projects is to save fresh water, facilitate water usage and distribution among stakeholders and farmers. To finance this project, conceptual cost models are important to accurately predict preliminary costs at the early stages of the project. FCIPs can be broke down into three main work elements; civil works, mechanical works, and electrical works. Civil works components include the construction of a pipeline, pump house, sump structure, suction pipes, and intake. Mechanical components include the installation of pump sets, irrigation valves, and mechanical connections. Electrical components include electrical boards and electrical connections (Radwan, 2013; Elmousalami et al., 2018b).

Elyamany and El-Nashar (2016) have presented an economic analysis and assessment for FCIPs taking into account the time value of money and Life Cycle Cost (LCC) methods. However, the previous study has not presented a model for cost prediction for FCIP. A total of 144 historical cases of FCIPs are randomly collected between 2010 and 2015 based on contracts information (Elmousalami et al., 2018b). Table 1 illustrates a descriptive statistics of collected data where mean and standard deviation are calculated for each variable where 17 variables are named from *P1* to *P17*.

5.2. *Quantitative key cost drivers ‘identification*

5.2.1. *Exploratory factor analysis*

Factor analysis is a ML method to convert correlated variables to a lower number of variables called factors. This method can be used to screen data to identify and categorize key parameters. EFA and confirmatory factor analysis (CFA) are two types of factor analysis (Polit and Beck, 2012). There are many types of factoring such as Principal Component Analysis (PCA), Canonical factor analysis and Image factoring etc. This study uses (PCA) which is a factor extraction method where factor weights are computed to extract the maximum possible variance. Subsequently, the factor model must be rotated for analysis (Polit and Beck, 2012). The advantage of EFA is to combine two or more variables into a single factor that reduces the number of variables. However, factor analysis cannot interpret the causality of the factored data. The following question should be answered to conduct EFA (Field, 2009):

(1) How large the sample needs to be?
(2) Is there multicollinearity or singularity?
(3) What is the method of data extraction?
(4) What is the number of factors to retain?
(5) What is the method of factor rotation?
(6) Choosing between factor analysis and principal components analysis?, all these questions will be answered in the following sections.

5.2.1.1. Sample size. Components are obtained from small datasets do not generalize as well as those derived from larger samples. Some

Table 1
Descriptive statistics.

ID	Variables	Unit	Minimum value	Maximum value
P_1	Area served	hectare	19	100
P_2	Average area served sections	hectare	2.65	13.1
P_3	Total length of pipeline	m	119	1832
P_4	Equivalent Diameter	mm	225	313.4
P_5	Duration (working days)	day	58	122.5
P_6	Irrigation valves number	unit	3	27
P_7	Air and pressure relief valves number	unit	1	7
P_8	sump (its diameter 1.7)	unit	0	1
P_9	Pump house (its size 3m*4m)	unit	0	1
P_{10}	Max discharge	liter/sec	40	120
P_{11}	Electrical pump discharge	liter/sec	40	120
P_{12}	Diesel pump discharge	liter/sec	40	120
P_{13}	Orientation	———	0	3
P_{14}	Construction year	year	2010	2015
P_{15}	Rice existence	———	0	1
P_{16}	Intake existence	unit	0	1
P_{17}	Ganabiaa canal	————	0	1

researchers have suggested using the ratio of sample size to the number of variables as a criterion. In the current study, there are *17* variables and the collected data set are *111* historical cases, the ratio between cases to variables is (6.5). According to (Kline, 1999) Guadagnoli and Velicer (1988), this ratio is initially acceptable. According to Tabachnick and Fidell (2007) and Comrey and Lee (1992), this sample size is classified as a poor sample. Furthermore, the Kaiser–Meyer–Olkin measure of sampling adequacy (KMO) (Kaiser, 1970).

Kaiser-Meyer-Olken)KMO(is another measure to compute the degree of inter-correlations among variables. The KMO statistic varies between zero and one (Kaiser, 1974). A value of zero shows that the sum of partial correlations is large relative to the sum of correlations, which means that there is diffusion in the pattern of correlations. Therefore, factor analysis is likely to be inappropriate. A value close to one shows that patterns of correlations are relatively compact and that factor analysis should yield reliable factors. In the present study, KMO measure of sampling adequacy is 0.69 which is classified as mediocre.

5.2.1.2. Correlation among variables. The first iteration to check correlation and avoid multicollinearity and singularity (Tabachnick and Fidell, 2007). Multicollinearity is that variables are correlated too highly whereas singularity is that variables are perfectly correlated. It is used to describe variables that are perfectly correlated (it means the correlation coefficient is 1 or -1). There are two methods for assessing multicollinearity or singularity:

1) The first method is conducted by scanning the correlation matrix for all independent variables to eliminate variables with correlation coefficients greater than 0.90 (Field, 2009) or correlation coefficients greater than 0.80 (Rockwell, 1975).

2) The second method is to scan the determinant of the correlation matrix (Heyrovsky, 1969). Multicollinearity or singularity may be in existence if the determinant of the correlation matrix is less than 0.00001. One simple heuristic is that the determinant of the correlation matrix should be greater than 0.00001 (Field, 2009). If the visual inspection reveals no substantial number of correlations greater than 0.3, PCA probably is not appropriate. Also, any variables that correlate with no others ($r = 0$) should be eliminated (Field, 2009).

The present study has conducted the following iterations:

Iteration 1: remove any variable higher than 0.9 with all independent variables to avoid singularity and multicollinearity (17 variables).
Iteration 2: the determinant of the correlation matrix is equal to 0.000 which is less than 0.00001. It implies that there is a problem with multicollinearity. By trial and error, it is found that *(P10) (P11) (P12)* and *(P13)* caused multicollinearity. Therefore, these factors have been deleted. Accordingly, the remaining variables are 13 variables.
Iteration 3: EFA is repeated for the third time after removing these parameters. The determinant of the correlation matrix is equal to 0.001, which is greater than 0.00001.
Iteration 4: the Anti-Image correlation matrix that contains Measures of Sampling Adequacy (MSA) is examined. All diagonal elements should be greater than 0.5 whereas the off-diagonal elements should all be very small (close to zero) in a good model (Field, 2009). The scan of Anti-image correlation matrix diagonal elements greater than 0.5 except three variables *P1, P15*, and *P16* which has values less than 0.5 equals to 0.459, 0.178, 0.383 respectively. Accordingly, the remaining variables have been reduced to ten variables.

5.2.1.3. Bartlett's test. Bartlett's test can be used to test the adequacy of the correlation matrix. It tests the null hypothesis that the correlation matrix is an identity matrix where all the diagonal values are equal to one and all off-diagonal values are equal to zero. A significant test indicates that the correlation matrix is not an identity matrix where a significance value less than 0.05 and null hypothesis can be rejected (Field, 2009). The significance value (*p-value*) = 0.000 where less than significance level. Therefore, it indicates that correlations between variables are sufficiently large for Factor Analysis.

5.2.1.4. Factor extraction by principal component analysis. Factor (component) extraction is the second step in conducting EFA to determine the smallest number of components that can be used to best represent interrelations among a set of variables (Tabachnick and Fidell, 2007). Communalities for retained variables after extraction are more than 0.5 which show that these variables are reflected well by extracted factors. Accordingly, the factor analysis is reliable (Field, 2009). The Kaiser criterion stated that if the number of variables is less than 30, then the average communality is more than 0.7 and if the number of variables is greater than 250, then the mean communality is near or greater than 0.6 (Stevens, 2002). Based on this criterion, only six parameters have been retained. The retained parameters are *P1, P3, P4, P5, P6*, and *P7*.

This result can be confirmed by retaining all components with eigenvalues more than 1 that contains five components. Table 2 illustrates initial eigenvalues with an eigenvalue of one or more are retained where that contains five components and percent of variance before and after the rotation. Table 3 illustrates the components with each parameter. Finally, Table 3 shows Rotated Component Matrix where the highest loading parameters for the first component are *P3, P1, P5, P4, P6,* and *P7* respectively.

5.2.2. Regression methods

Regression analysis can be used for both cost drivers selection and cost prediction modeling. The current study focused on the cost driver's selection. Therefore, the forward, backward, stepwise methods are applied as follows.

5.2.2.1. Forward method. Forward selection initiates with no variables in the model where each added variable is tested by a comparison criterion to improve the model performance. If the independent variable

Table 2
Total variance explained.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings	
	Total	% of Variance	Cumulative %	Total	% of Variance
1	4.628	35.601	35.601	4.628	35.601
2	1.545	11.887	47.488	1.545	11.887
3	1.307	10.051	57.539	1.307	10.051
4	1.083	8.331	65.870	1.083	8.331
5	1.006	7.735	73.605	1.006	7.735

Table 3
Component matrix.

ID	Component				
	1	2	3	4	5
P3	.847		-.197		.292
P1	.844	.219	.123		-.213
P5	.821	-.228	.135	.361	.241
P4	.737	.355			-.315
P6	.728	-.248	-.217	-.128	-.201
P7	.565	.110	-.424		.561

significantly improves the ability of the model to predict the dependent variable, then this predictor is retained in the model and the computer searches for a second independent variable (Field, 2009). Results are illustrated in Table 4 where the model (1) included a single variable (*P3*) and its correlation factor is (0.85).

5.2.2.2. *Backward elimination method.* The backward method is the opposite of the forward method. In this method, all input independent variables are initially selected, and then the most unimportant independent variable are eliminated one-by-one based on the significance value of the t-test for each variable. The contribution of the remaining variable is then reassessed (Field, 2009). The results are illustrated in Table 5 where the model (1) includes 16 variables and its correlation factor is (0.96).

5.2.2.3. *Stepwise method.* Stepwise selection is an extension of the forward selection approach, where input variables may be removed at any subsequent iteration (Field, 2009). Despite forward selection, Stepwise selection tests at each step for variables to be included or excluded where stepwise is a combination of backward and forward methods. The results are illustrated in Table 6 where the model (1) includes a single variable (*P3*) and its correlation factor is (0.85).

5.2.3. *Correlation*

The relation among all variables are shown in the correlation matrix, the aim is to screen variable based only on the correlation matrix. Therefore, all independent variables that are highly correlated with each other will be eliminated (*R*>= *0.8*) and all dependent variables that are low correlated with the dependent variable (*R* =*0.3*).

Pearson Correlation is a measure of the linear correlation between two variables, giving a value between +1 and −1 where 1 is the positive correlation, 0 is no correlation, and −1 is the negative correlation. It is developed by Karl Pearson as a measure of the degree of linear dependence between two variables (Field, 2009). Correlation(r) can computed based on Eq. (1).

$$r = \frac{\sum XY - n\overline{X}\overline{Y}}{\sqrt{\sum X^2 - n\overline{X}^2}\sqrt{\sum Y^2 - n\overline{Y}^2}} \tag{1}$$

where.

r: the correlation coefficient.
X: dependent variable.
Y: Independent variable.
n: number of data points.
X: the mean of *X* values.
Y: the mean of *Y* values.

After the first scan of the correlation matrix, it found that: First, the correlation among independent variables *(P10) (P11) (P12)* and *(P13)* is highly correlated with *(P1)* where the correlation factor is approximately 0.86 for them. Second, the correlation among independent variables and the dependent variable *(P14), Rice (P15) (P16)* and *(P17)* are low correlated with the dependent variable (the cost of FCIP) where there is no relation between them. The correlation coefficient are 0.071, 0.206, 0.036, 0.104 and 0.19 respectively. Therefore, these variables are eliminated and correlation matrix scanned for the second iteration where all correlations are significant at *P=0.01* (level 2-tailed). The selected cost drivers are *P1, P3, P4, P5, P6, P7, P8* and *P9.*

Table 4
Forward method results.

Model	Independent Variable	R	R Square	Adjusted R Square
1	*(P3)*	0.85	0.73	0.72
2	*(P3), (P14)*	0.89	0.80	0.79
3	*(P3), (P14), (P10)*	0.92	0.84	0.84
4	*(P3), (P14), (P10), (P11)*	0.93	0.87	0.86
5	*(P3), (P14), (P10), (P11), (P9)*	0.94	0.89	0.89
6	*(P3), (P14), (P10), (P11), (P9), (P5)*	0.95	0.90	0.90
7	*(P3), (P14), (P10), (P11), (P9), (P5), (P8)*	0.95	0.91	0.90
8	*(P3), (P14), (P10), (P11), (P9), (P5), (P8), (P6)*	0.96	0.92	0.91

Table 5
Backward elimination method results.

Model	Independent Variable	R	R Square	Adjusted R Square
1	*(P17), (P7), (P14), (P15), (P2), (P16), (P13), (P8), (P9), (P6), (P4), (P3), (P12), (P1), (P5), (P10)*	0.96	0.93	0.92
2	*(P17), (P7), (P14), (P15), (P2), (P16), (P13), (P8), (P9), (P6), (P3), (P12), (P1), (P5), (P10)*	0.96	0.93	0.92
3	*(P17), (P7), (P14), (P15), (P2), (P13), (P8), (P9), (P6), (P3), (P12), (P1), (P5), (P10)*	0.96	0.93	0.92
4	*(P17), (P7), (P14), (P15), (P2), (P8), (P9), (P6), (P3), (P12), (P1), (P5),(P10)*	0.96	0.93	0.92
5	*(P17), (P7), (P14), (P15), (P2), (P8), (P9), (P6), (P3), (P12), (P5), (P10)*	0.96	0.93	0.92
6	*(P17), (P7), (P14), (P15), (P2), (P8), (P9), (P6), (P3), (P12), (P10)*	0.96	0.93	0.92
7	*(P7), (P14), (P15), (P2), (P8), (P9), (P6), (P3), (P12), (P10)*	0.96	0.93	0.92

Table 6
Stepwise Method results.

Model	Independent Variable	R	R Square	Adjusted R Square
1	*(P3)*	0.85	0.73	0.72
2	*(P3), (P14)*	0.89	0.80	0.79
3	*(P3), (P14), (P10)*	0.92	0.84	0.84
4	*(P3), (P14), (P10), (P11)*	0.93	0.87	0.86
5	*(P3), (P14), (P10), (P11), (P9)*	0.94	0.89	0.89
6	*(P3), (P14), (P10), (P11), (P9), (P5)*	0.95	0.90	0.90
7	*(P3), (P14), (P10), (P11), (P9), (P5), (P8)*	0.95	0.91	0.90
8	*(P3), (P14), (P10), (P11), (P9), (P5), (P8), (P6)*	0.96	0.92	0.91

Spearman correlation is a nonparametric measure of statistical dependence between two variables using a monotonic function. A perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other (Field, 2009). Spearman correlation is more general than Pearson's coefficient, which only measures linear dependence. First, the Correlation among independent variables (P10) (P11) and (P12) is highly correlated with (P1) where correlation factor is approximately 0.86 for them. Second, Correlation among independent variables and the dependent variable (P14) (P15), and (P16) and (P17) are low correlated with the dependent variable (the cost of FCIP) where there is no relation between them. The correlation coefficient are

0.12, 0.18, 0.05, 0.14 and 0.19 respectively. Therefore, these variables are eliminated and correlation matrix scanned for the second time, all correlations are significant at P = 0.01 (level 2-tailed). The selected variables are (P1) (P3) (P4) (P5) (P6) (P7) (P8) and (P9) and (P13) as shown in Table.9 and Fig. 5.

5.2.4. Hybrid feature models

Many limitations exist to the previous methods. According to the EFA model, this model needs a sufficient data set to be successfully conducted. Furthermore, if observed variables are highly similar to each other, factor analysis will identify a single factor to them. According to regression-based models, statistically, there are several points of criticism. Wilkinson and Dallal (1981) indicated that testes are biased where there is a difference in significance level in the F-procedure as a test of forward regression method. To avoid model overfitting model, the expert judgment may be needed to validate the selected variables and the model instead of the validation data set (Flom and Cassell, 2007). According to correlation models, correlation cannot imply causation where a correlation between two variables is not a condition to identify a causal relationship. A correlation coefficient is not sufficient to identify the dependence structure between random variables (Mahdavi Damghani B., 2013). The strength of a linear relationship between two variables can be identified by the Pearson correlation coefficient. However, its value generally does not completely characterize their relationship (Mahdavi Damghani, Babak, 2012).

ML models can be conducted subsequently to produce better performance (Bishop, 2006). A hybrid model is to merge two methods as one method to obtain better results as illustrated in Fig. 4. The first model is to conduct Pearson correlation where all independent variables are highly correlated ($R>=0.8$) is eliminated and all independent variables low correlated ($R=0.3$) with dependent variable are eliminated, and then to conduct stepwise method to identify the final selection of variables. The results are shown in Table 7. The second hybrid model is to conduct the first approach without deleting independent variables low correlated ($R = 0.3$) with the dependent variable as shown in Table 8.

5.2.5. Discussion of key cost drivers' results

In the current study, the correlation coefficient between the dependent variable and the independent variable is used as the benchmark to compare the results of variable extraction methods. Table 9 and Fig. 5 summarize the results of methods where correlation is shown against the number of variables. Fig. 5 can be used by a model developer to choose key cost drivers based on the following two criteria: First, the fewer number of variables that can represent the highest correlation with the outcome (cost of FCIP). Second, the availability of data at the conceptual stage where this chart provides a set of alternatives of the selected variables (cost drivers) to give the same accuracy with the outcome. For example, if the model developer wants to develop a model with the highest accuracy, Fig. 5 suggests using backward elimination method to provide high correlation ($R=0.96$). However, the number of the required variables may be ten variables and that may not be available at the conceptual stage of the project.

The second logic trial is to use the fewer variables (assume four variables), Fig. 5 suggests the two methods with approximately the same correlation (Stepwise and Hybrid model 2). By looking at the corresponding Table 6 and Table 8, Stepwise method variables are *(P3) (P14) (P10), and (P11)* whereas Hybrid model (2) is *(P3) (P14) (P6),* and *(P1)*. At this phase the model developer has two options to develop the proposed model, the choice will depend on the second criteria (availability of data at the conceptual stage). The final selection is based on the hybrid model (2). Accordingly, the four key cost drivers are *(P3) (P14) (P6)*, and *(P1)*.

For validation purpose, Elmousalami et al. (2018a) have conducted a study to identify key cost drivers of FCIPs based on only qualitative techniques such as Delphi method, fuzzy Delphi techniques, and fuzzy analytical hierarchy process. Whereas, the current study has identified the key cost drivers based on the quantitative statistical techniques (based on the collected data, not experts) such as Pearson correlation and stepwise regression. Elmousalami et al. (2018a) have identified the key cost drivers for FCIPs based on qualitative techniques. The final key cost drivers are area served (P1), pipeline total length (P3), year of construction (P14) and irrigation valves number (P6). As a result, the identified cost drivers based on qualitative methods are matching with the resulting cost drivers based on the second hybrid model. The key advantage of the hybrid model than qualitative methods is the automation of key cost drivers' identification where no experts have been needed.

5.3. Fuzzy modeling

5.3.1. Application and analysis of traditional fuzzy model

The objective is to develop a parametric cost estimate model for FCIP for conceptual feasibility studies and cost estimation purposes. The collected data is divided into two sets: training set (89 cases, 80%) and validation set (22 cases, 20%). Once the key cost drivers are identified based on the quantitative approaches, these cost drivers can be applied as inputs to the fuzzy model. Therefore, the following step is to fuzzification the four key cost drivers and identify their MFs as shown in Fig. 6. The most critical stage is to develop fuzzy rules base. Traditionally, experts are consulted to give their experience to develop such rules. For example, the current case study consists of four key cost drivers, each cost driver consists of seven MFs as shown in Fig. 6. For example, the input variable construction year consists of seven triangle MFs {MF_1, MF_2, MF_3, MF_4, MF_5, MF_6, MF_7} as shown in Fig. 6. Accordingly, the number of possible rules may equal ($7^4 = 2401$ rules). Therefore, there is a need to automatically generate such rules.

5.3.2. Application and analysis of genetic-fuzzy model

On the other hand, Genetic-Fuzzy model has been developed to optimally generate fuzzy rules. The study has applied GA to optimally select the fuzzy rules where 2401 rules represent the whole possible search space for GA. The formulation of genetic algorithm model depends mainly on defining two core terms: a chromosome representation and an objective function. First, based on Michigan approach, the chromosomes represents the fuzzy rules where the number of chromosomes (CH_n) are the number of fuzzy rules. Each chromosome is consists of five genes where four genes for the key input parameters *(P_1, P_2, P_3, P_4)* respectively and the fifth gene is for the output (the cost of FCIP). Each gene consists of one of the seven membership functions (MF_i) where (i) is ranging from one to seven *(MF_1:MF_7)* as shown in Fig. 7. For example: **IF** {Area served (P_1) is MF_5 **AND** Total length (P_2) is MF_2 **AND** Irrigation valves (P_3) is MF_2 **AND** construction year (P_4) is MF_6} **THEN** {The Cost LE/Mesqa is MF_3}.

Secondly, the fitness function is problem-dependent where the objective is to enhance the accuracy and quality of the system performance (Hatanaka et al., 2004). The fitness function can be formulated as

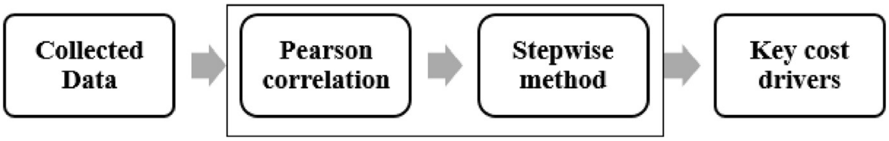


Fig. 4. The hybrid ML model concept.

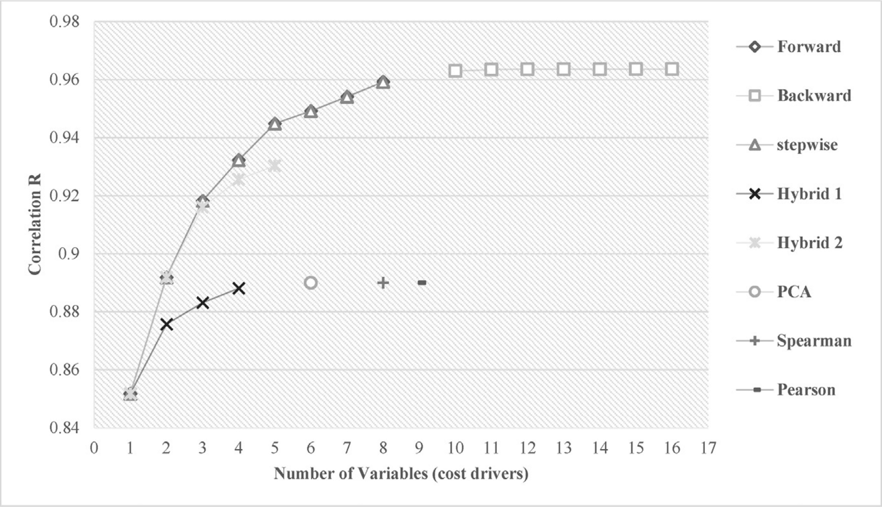


Fig. 5. All results are plotted for each method.

Table 7
The results of the first iteration of the hybrid model (1).

Hybrid Model (1)	Independent Variable	R	R Square	Adjusted R Square
1	*(P3)*	0.85	0.73	0.72
2	*(P3), (P1)*	0.88	0.77	0.76
3	*(P3), (P1), (P6)*	0.88	0.78	0.77
4	*(P3), (P1), (P6), (P7)*	0.89	0.79	0.78

Table 8
The results of the first iteration of the hybrid model (2).

Hybrid Model (2)	Independent Variable	R	R Square	Adjusted R Square
1	*(P3)*	0.85	0.73	0.72
2	*(P3), (P14)*	0.89	0.80	0.79
3	*(P3), (P14), (P6)*	0.92	0.84	0.83
4	*(P3), (P14), (P6), (P1)*	0.93	0.86	0.85
5	*(P3), (P14), (P6), (P1), (P9)*	0.93	0.87	0.86

Table 9
Results of all methods.

Method	Select Variables	R	Number of variables
EFA	*(P3), (P1), (P5), (P4),(P6) and (P7)*	0.89	6
Forward Method	Table 4		
Backward Method	Table 5		
Stepwise Method	Table 6		
Pearson Correlation	*(P1), (P3), (P4), (P5), (P6), (P7), (P8) and (P9)*	0.89	8
Spearman Correlation	*(P1), (P3), (P4), (P5), (P6), (P7), (P8) and (P9) and (P13)*	0.89	9
hybrid model (1)	Table 7		
hybrid model (2)	Table 8		

the following Eq. (2) where the objective is to maximize the fitness function and minimize MAPE and the number of rules.

$$\boldsymbol{Max}\ (\boldsymbol{F}) = \left(\frac{1}{\boldsymbol{MAPE} + \boldsymbol{Nr}}\right) \tag{2}$$

where: (F) is a fitness function and (Nr) is the number of rules and MAPE is as following Eq. (3):

$$\boldsymbol{MAPE} = \frac{1}{\boldsymbol{n}}\sum_{i=1}^{n}\frac{\left|(\boldsymbol{Y}_{actual})_i - (\boldsymbol{Y}_{model})_i\right|}{(\boldsymbol{Y}_{model})_i}\boldsymbol{x}100 \tag{3}$$

where (n) is the number of cases, (i) is the number of the case and Y_{model} is the outcome of model and Y_{actual} is the actual outcome. Moreover, MAPE can be replaced by the mean square error (MSE) as following Eq. (4):

$$\boldsymbol{MSE} = \frac{1}{\boldsymbol{n}}\sum_{i=1}^{n}\left[(\boldsymbol{Y}_{actual})_i - (\boldsymbol{Y}_{model})_i\right]^2 \tag{4}$$

As shown in Fig. 7, the process of the developed model consists of five main steps:

1. An initial population of chromosomes has been identified to represent the initial state of the fuzzy rules. The four key cost drivers have been fed to the fuzzy system.
2. The fuzzy system produces the final output of the system (cost of FCIP) Y $_{model}$
3. The Y_{model} has been fed to fitness function (F) to evaluate the model performance.
4. GA uses the fitness function (F) to evaluate the search process where crossover probability and mutation probability have been set at 0.7 and 0.01 respectively.
5. The new population of fuzzy rules has been produced based on crossover and mutation processes.

5.3.3. Discussion of fuzzy models' results

As shown in Table.10, the number of generated rules by the GA-fuzzy model were 63 rules and the MAPE was 14.7%. On the other hand, a traditional fuzzy model has been built based on the experts ‘experience where a total of 190 rules were generated to cover all the possible combinations of the fuzzy system and MAPE was 26.3 %. According to the cost of computation, the traditional fuzzy model needs computation cost less than the genetic-fuzzy model. Moreover, the results show that the rules generated by experts may have redundant rules which can be deleted to improve the model computation and performance. Moreover, the expert's knowledge cannot cover all combination to represent all possible rules (2401 rules). In addition, the generation of the experts' rules is time and effort consuming process. However, the GA approach provides fewer rules that optimally cover all the possible rules and

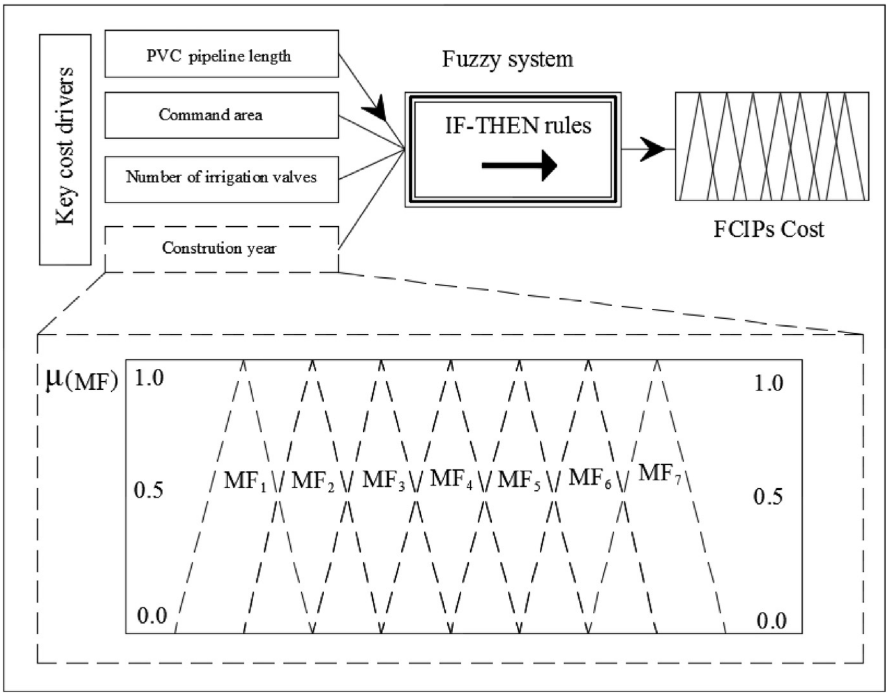
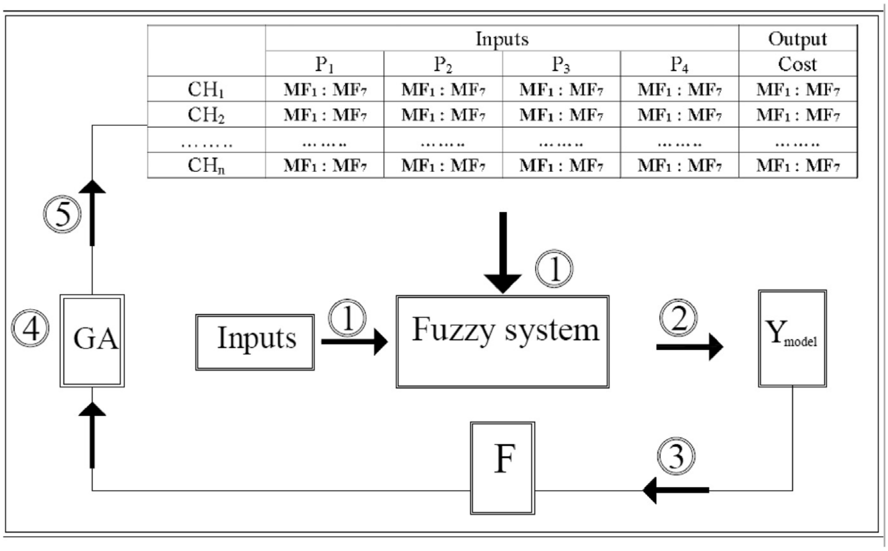


Fig. 6. Fuzzy system for FCIPs, and MFs for key cost drivers.



	Inputs				Output
	P_1	P_2	P_3	P_4	Cost
CH_1	$MF_1 : MF_7$	$MF_1 : MF_7$	$MF_1 : MF_7$	$MF_1 : MF_7$	$MF_1 : MF_7$
CH_2	$MF_1 : MF_7$	$MF_1 : MF_7$	$MF_1 : MF_7$	$MF_1 : MF_7$	$MF_1 : MF_7$
………	……..	……..	……..	……..	……..
CH_n	$MF_1 : MF_7$	$MF_1 : MF_7$	$MF_1 : MF_7$	$MF_1 : MF_7$	$MF_1 : MF_7$

Fig. 7. The process of genetic fuzzy system.

provide the optimal accuracy and performance of the developed system.

For model validation, Elmousalami et al. (2018b) have developed a quadratic regression model and ANNs that can predict the conceptual cost for FCIPs at 9% MAPE. Data transformation plays an important role in prediction accuracy. However, the main gap of this research is lacking the uncertainty modeling to the prediction cost model. Although the prediction accuracy of the quadratic regression modeling was approximately 9% better than the genetic fuzzy model by approximately 5.7%, the fuzzy model would produce more reliable prediction results due to taking uncertainty into account. Fig. 8 shows the actual and predicted cost for fuzzy and genetic fuzzy models where the genetic fuzzy model has improved the performance from $R^2 = 0.61$ to $R^2 = 0.77$. This improvement results from a good representation of fuzzy rules using genetic algorithm optimization.

The proposed intelligent methodology consists of two main stages. The first stage is feature engineering using the collected data to automatically identify the key cost drivers. The select cost drives provides the

Table 10
Comparison of traditional fuzzy model and genetic-fuzzy model.

Criterion\model	Traditional fuzzy model	Genetic-fuzzy model
MAPE	26.3%	14.7%
R^2	0.61	0.77
Generated fuzzy rules	190 of 2401	63 of 2401
Fuzzy rules generation	By experts	Automated by GA
Computation complexity	Less than Genetic fuzzy model	More than the traditional fuzzy model
Time and effort	High	Low

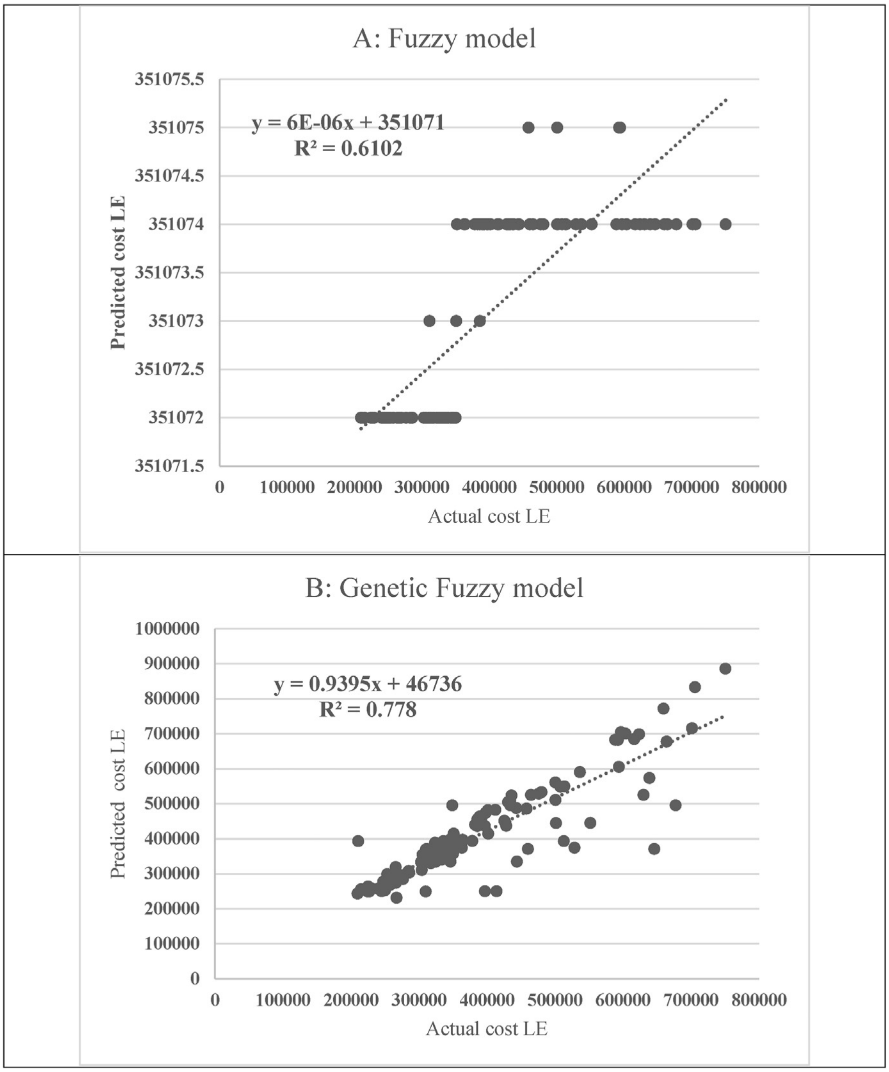


Fig. 8. (A) Actual and predicted cost for fuzzy logic system (B) Actual and predicted cost for genetic fuzzy system.

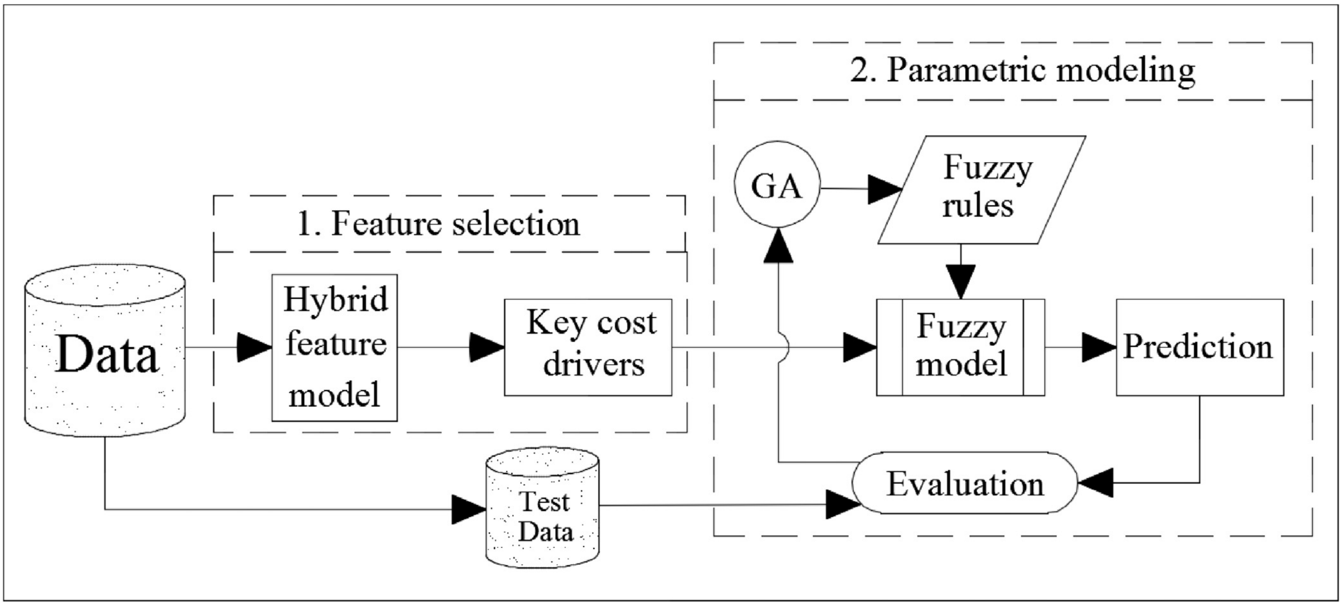


Fig. 9. Intelligent methodology architecture.

high accuracy based on the least number of predictors as shown in Fig. 5 where hybrid model 2 has been select to filter the given data. The second stage is parametric modeling using the data to predict the conceptual cost with considering uncertainty and achieving the most accurate results. Genetic fuzzy model has been develop to formulate cost estimating relationships (CERs) taking uncertainty into account. These two stages have been incorporated together to build the intelligent methodology for conceptual cost prediction as shown in Fig. 9.

The following points summarize the recommendations and future trends:

I. This study recommended applying both qualitative (Elmousalami et al., 2018a) and quantities approache to obtain the most reliable cost drives. Such a procedure can be called a hybrid approach for cost drivers' identification. The limitation of the hybrid approach is that both experts' and historical cases are required to be operated on.
II. The conceptual cost estimate is conducted under uncertainty. Therefore, this study recommended using fuzzy theory such as Fuzzy Logic and to develop a hybrid model based on fuzzy logic to obtain uncertainty nature for the developed model and produce more reliable performance.
III. The hybrid fuzzy model presents a superior performance than the single fuzzy model. Fuzzy membership function and fuzzy rules can be automated based on hybrid fuzzy models to produce more reliable predictions. Therefore, this study recommends developing an automated hybrid fuzzy rules models than traditional fuzzy models. In addition, this recommendation can be generalized not only for fuzzy cost estimation models but also for all fuzzy modeling in different applications. Accordingly, hybrid fuzzy modeling is a future research trend in engineering prediction and computation modeling.
IV. The Genetic algorithm is a powerful tool to select the optimal set of the cost drivers where the prediction error is minimized. GAs is used widely for hybrid model development.
V. Computational models and information systems have been applied in business and construction industry to effectively improve the job efficiency (Davis, 1993). Therefore, the hybrid model represents the current trend of parametric cost modeling to improve the model performance and accuracy where the limitations of each technique can be avoided. The objective is to develop computerized automated systems with less interventions of humans to save time, effort and avoid human error for cost estimate. Moreover, computer technologies have a great ability to deal with vast data and complicated computations.
VI. There is a need to develop a model that has the ability to give justification on the model's results and to give answers and interpretations for the predicted cost. That may require a higher level of AI and may represent the future trend of cost modeling. Moreover, such a concept may be generalized for any prediction model. The objective is to avoid the estimator's biases, warn the user to the input parameters of the model, and to avoid the limitation of the black box nature.

6. Conclusion

The study presents an intelligent sequential methodology which provides a roadmap for conceptual cost modeling. The main contribution of the methodology is to avoid depending on experts' opinions to identify the key cost drivers and to take the uncertainty concept in the cost prediction modeling. The methodology automatically determines the cost drivers based only on the quantitative data without using on experts' opinions such as questionnaire survey, Delphi methods, or analytical hierarchy process. The methodology conducts fuzzy logic theory to prediction modeling to maintain uncertainty to the prediction value. The present study has discussed fuzzy modeling and its benefit to obtain uncertainty for the studied case. In addition, the study highlights the main problem for fuzzy modeling which is fuzzy rules generation. This study has reviewed the hybrid fuzzy model methodologies to generate rules such as evolutionary fuzzy model. Moreover, a case study has been conducted to investigate the effectiveness of quantitative approaches than qualitative approach and hybrid fuzzy modeling such as genetic-fuzzy model than traditional fuzzy modeling. The study emphasizes the importance of quantitative approaches for key cost drives and the genetic fuzzy model for conceptual cost prediction.

The methodology has been validated by application to FCIPs where this methodology have been compared to a tradition methodology based on qualitative cost drivers' identification (Elmousalami et al., 2018a) and deterministic cost prediction (Elmousalami et al., 2018b). However, many limitations exist in the study:

1 The methodology needs to be applied for several projects to be generalized.
2 Applying correlation methods to select the key cost drivers may produce wrong cost drivers' identification as the correlation doesn't mean causation. Therefore, it is recommended to apply both qualitative and quantitative approaches to obtain the most reliable cost drives. Such a procedure can be called a hybrid approach for cost drivers' identification. The limitation of the hybrid approach is that both experts' and historical cases are required to be operated on.
3 Applying fuzzy logic model may decrease the prediction accuracy where regression and ANNs may produce a better prediction accuracy. However, applying fuzzy theory is recommended to take uncertainty concepts to the conceptual cost model.
4 Automation the cost models are prone to many machine learning problems such as overfitting issues, hyper-parameter selection, and validation.
5 This study has not discussed all models such as neuro-fuzzy models and MF generation. Therefore, further studies needed to compare and analyze the different performances of such models.

Therefore, several future research studies are needed to develop these limitations.

Declarations

Author contribution statement

Haytham ElMousalami: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

This study presents the publicly open dataset FCIPs to be used for validation and analysis of the future cost prediction algorithms. Dataset available at https://github.com/HaythamElmousalami/Field-canals-improvement-projects-FCIPs-

References

AACE International recommended practices, 2004. AACE International. Morgantown, W.V.

Ahiaga-Dagbui, D.D., Tokede, O., Smith, S.D., Wamuziri, S., 2013. A neuro-fuzzy hybrid model for predicting final cost of water infrastructure projects. In: Procs 29th Annual ARCOM Conference, 2-4 September 2013. Association of Researchers in Construction Management, Reading, UK, pp. 181–190.

Alroomi, A., Jeong, D.H.S., Oberlender, G.D., 2012. Analysis of cost-estimating competencies using criticality matrix and factor Analysis. J. Constr. Eng. Manage. 138 (11), 1270–1280.

Ambrule, V.R., Bhirud, A.N., 2017. Use of artificial neural network for pre design cost estimation of building projects. Int. J Recent Innovat. Trends Comput. Commun. 5 (2), 173–176.

Angelov, P.P., 2002. Evolving rule-based models. A Tool for Design of Flexible Adaptive Systems. Physica-Verlag, Wurzburg.

Arabzadeh, V., Niaki, S.T.A., Arabzadeh, 2018. Construction cost estimation of spherical storage tanks: artificial neural networks and hybrid regression—GA algorithms" V. J Ind Eng Int 14, 747.

Attalla, M., Hegazy, T., 2003. Predicting cost deviation in reconstruction projects: artificial neural networks versus regression. J. Constr. Eng. Manag. 129 (4), 405–411.

Belarbi, K., Titel, F., 2000. Genetic algorithm for design of a class of fuzzy controllers: an alternative approach. IEEE Trans. Fuzzy Syst. 8 (4), 398–405.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer Verlag, New York, NY.

Bonarini, A., 1996. Evolutionary learning of fuzzy rules: competition and cooperation. Fuzzy Modelling. Springer, Boston, MA, pp. 265–283.

Bonarini, A., Trianni, V., 2001. Learning fuzzy classifier systems for multi-agent coordination. Inf. Sci. 136, 215–239.

Cao, Y., Ashuri, B., Baek, M., 2018. Prediction of unit price bids of resurfacing highway projects through ensemble machine learning. J. Comput. Civ. Eng. 32 (5), 04018043.

Casillas, J., Carse, B., Bull, L., 2007. Fuzzy-xcs: a Michigan genetic fuzzy system. IEEE Trans. Fuzzy Syst. 15 (4), 536–550.

Cheng, M.-Y., Roy, A.F., 2010. Evolutionary fuzzy decision model for construction management using support vector machine. Expert Syst. Appl. 37 (8), 6061–6069.

Cheng, M.-Y., Tsai, H.-C., Hsieh, W.-S., 2009. Web-based conceptual cost estimates for construction projects using Evolutionary Fuzzy Neural Inference Model. Autom. ConStruct. 18 (2), 164–172.

Cheng, M.Y., Tsai, H.C., Sudjono, E., 2010 Jun 1. Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. Expert Syst. Appl. 37 (6), 4224–4231.

Chou, C.-H., 2006. Genetic algorithm-based optimal fuzzy controller design in the linguistic space. IEEE Trans. Fuzzy Syst. 14 (3), 372–385.

Comrey, A.L., Lee, H.B., 1992. A First Course in Factor Analysis, second ed., 2. Erlbaum Conf. on Fuzzy Systems (FUZZ-IEEE'94), Hillsdale, NJ, pp. 1377–1382. Orlando, FL, USA, 1994.

Cordon, O., Herrera, F., Hoffmann, F., Magdalena, L., 2001. Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases. World Scientific, Singapore.

Damghani, B.M., Welch, D., O'Malley, C., Knights, S., 2012 Nov. The misleading value of measured correlation. Wilmott 2012 (62), 64–73.

Davis, F.D., 1993. User acceptance of information technology: system characteristics, user perceptions, and behavioral impacts. Int. J. Man Mach. Stud. 38, 475–487.

Dell'Isola, M.D., 2002. Architect's Essentials of Cost Management, 8. John Wiley & Sons.

Dursun, O., Stoy, C., 2016. Conceptual estimation of construction costs using the multistep ahead approach. J. Constr. Eng. Manag. 142 (9), 04016038.

Elmousalami, H.H., Elyamany, A.H., Ibrahim, A.H., 2018a Jan 1. "Evaluation of Cost Drivers for Field Canals Improvement Projects. Water Resour. Manag. 32 (1), 53–65.

Elmousalami, H.H., Elyamany, A.H., Ibrahim, A.H., 2018b. "Predicting conceptual cost for field canal improvement projects. J. Constr. Eng. Manag. 144 (11), 04018102.

ElSawy, I., Hosny, H., Abdel Razek, M., 2011. A neural network model for construction projects site Overhead cost estimating in Egypt. IJCSI Int. J. Comput. Sci. Issues 8 (3). No. 1, May 2011 ISSN (Online): 1694-0814.

Elyamany, A.H., El-Nashar, W.Y., 2016. Estimating life cycle cost of improved field irrigation canal. Water Resour. Manag. 30 (1), 99–113.

Field, A., 2009. Discovering Statistics Using SPSS for Windows. Sage Publications, London e Thousand Oaks e New Delhi.

Flom, P.L., Cassell, D.L., 2007 Nov 11. Stopping stepwise: why stepwise and similar selection methods are bad, and what you should use. In: North East SAS Users Group Inc 20th Annual Conference: 11-14th November 2007; Baltimore, Maryland.

Goldberg, D.E., 2002. The Design of Competent Genetic Algorithms: Steps toward a Computational Theory of Innovation.

Green, S.B., 1991. How many subjects does it take to do a regression analysis? Multivariate Behav. Res. 26, 499–510.

Guadagnoli, E., Velicer, W.F., 1988. Relation of sample size to the stability of component patterns. Psychol. Bull. 103 (2), 265–275.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Hatanaka, T., Kawaguchi, Y., Uosaki, K., 2004. Nonlinear system identification based on evolutionary fuzzy modelling. IEEE Congr. Evol. Comput. 1, 646–651.

Hegazy, T., Ayed, A., 1998. Neural network model for parametric cost estimation of highway projects. J. Constr. Eng. Manage. 124 (3), 210–218.

Herrera, F., 2008. Genetic fuzzy systems: taxonomy, current research trends and prospects. Evol. Intell. 1 (1), 27–46.

Herrera, F., Verdegay, J.L. (Eds.), 1996. Genetic Algorithms and Soft Computing. Physica-Verlag, Wurzburg, pp. 152–171.

Heyrovsky, Y., 1969. Multicollinearity in regression analysis: a comment. Rev. Econ. Stat. 51 (4), 486–489.

Hinton, G.E., Nowlan, S.J., 1987. How Learning Can Guide Evolution, Complex Systems, pp. 495–502.

Hoffmann, F., Pfister, G., 1997. Evolutionary design of a fuzzy knowledge base for a mobile robot. Int. J. Approx. Reason. 17 (4), 447–469.

Homaifar, A., McCormick, E., 1995. Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms. IEEE Trans. Fuzzy Syst. 3 (2), 129–139.

Hsu, Y.-L., Lee, C.-H., Kreng, V., 2010. The application of Fuzzy Delphi Method and Fuzzy AHP in lubricant regenerative technology selection. Expert Syst. Appl. 37 (1), 419–425.

Ishibuchi, H., Nakashima, T., Murata, T., 1999. Performance evaluation of fuzzy classi5er systems for multidimensional pattern classi5cation problems. IEEE Trans. System Man Cybernet. 29, 601–618.

Jin, R., Cho, K., Hyun, C., Son, M., 2012 Apr 1. MRA-based revised CBR model for cost prediction in the early stage of construction projects. Expert Syst. Appl. 39 (5), 5214–5222.

Juszczyk, M., 2017. Studies on the ANN implementation in the macro BIM cost analyzes. Przegląd Naukowy. Inżynieria i Kształtowanie Środowiska 26 (2), 76.

Juszczyk, M., Leśniak, A., Zima, K., 2018. ANN Based Approach for Estimation of Construction Costs of Sports fields. Complexity.

Kaiser, H.F., 1970. A second generation little jiffy. Psychometrika 35 (4), 401–415.

Kaiser, H.F., 1974. An index of factorial simplicity. Psychometrika 39 (1), 31–36.

Karatas, Y., Ince, F., 2016. Feature article: fuzzy expert tool for small satellite cost estimation. IEEE Aerosp. Electron. Syst. Mag. 31 (5), 28–35.

Karr, C.L., Gentry, E.J., 1993. Fuzzy control of pH using genetic algorithms. IEEE Trans. Fuzzy Syst. 1 (1), 46–53.

Kim, G.-H., Shin, J.-M., Kim, S., Shin, Y., 2013. Comparison of school building construction costs estimation methods using regression analysis, neural network, and support vector machine. J. Build. Constr. Plan. Res. 01 (01), 1–7.

Kim, S., Chin, S., Kwon, S., 2019. A discrepancy analysis of BIM-based quantity take-off for building interior components. J. Manag. Eng. 35 (3), 05019001.

Kline, P., 1999. The Handbook of Psychological Testing, second ed. Routledge, London.

Kwon, C., Sudhoff, S.D., 2006. Genetic algorithm-based induction machine characterization procedure with application to maximum torque per amp control. IEEE Trans. Energy Convers. 21 (2), 405–415.

Leśniak, A., Zima, K., 2018. Cost calculation of construction projects including sustainability factors using the Case Based Reasoning (CBR) method. Sustainability 10 (5), 1608.

Linkens, D.A., Nyongesa, H.O., 1995a. Genetic algorithms for fuzzy control, Part 1: offline system development and application. IEE Proc. Control Theory Appl. 142 (3), 161–176.

Linkens, D.A., Nyongesa, H.O., 1995b. Genetic algorithms for fuzzy control, Part 2: online system development and application. IEE Proc. Control Theory Appl. 142 (3), 177–185.

Liu, W.-K., 2013. Application of the Fuzzy Delphi Method and the Fuzzy Analytic Hierarchy Process for the Managerial Competence of Multinational Corporation Executives.

Loop, B.P., Sudhoff, S.D., Zak, S.H., Zivi, E.L., 2010. Estimating regions of asymptotic stability of power electronics systems using genetic algorithms. IEEE Trans. Control Syst. Technol. 18 (5), 1011–1022.

Lowe, D.J., Emsley, M.W., Harding, A., 2006. Predicting construction cost using multiple regression techniques. J. Constr. Eng. Manage. 132 (7), 750–758.

Mahdavi Damghani, B., 2013. The non-misleading value of inferred correlation: an introduction to the cointelation model. Wilmott Magazine.

Manoliadis, O.G., Pantouvakis, J.P., Christodoulou, S.E., 2009. "Improving qualifications-based selection by use of the fuzzy Delphi method. Constr. Manag. Econ. 27 (4), 373–384.

Marzouk, M.M., Ahmed, R.M., 2011. A case-based reasoning approach for estimating the costs of pump station projects. J. Adv. Res. 2 (4), 289–295.

Marzouk, M., Alaraby, M., 2014. Predicting telecommunication tower costs using fuzzy subtractive clustering. J. Civ. Eng. Manag. 21 (1), 67–74.

Marzouk, M., Amin, A., 2013. Predicting construction materials prices using fuzzy logic and neural networks. J. Constr. Eng. Manag. 139 (9), 1190–1198.

Marzouk, M., Elkadi, M., 2016. Estimating water treatment plants costs using factor analysis and artificial neural networks. J. Clean. Prod. 112, 4540–4549.

Ogungbile, A.J., Oke, A.E., Rasak, K., 2018. Developing cost model for preliminary estimate of road projects in Nigeria. International Journal of Sustainable Real Estate and Construction Economics 1 (2), 182–199.

Pedrycz, W. (Ed.), 1997. Fuzzy Evolutionary Computation. Kluwer Academic Publishers", Dordrecht.

Petroutsatou, K., Georgopoulos, E., Lambropoulos, S., Pantouvakis, J.P., 2012. Early cost estimating of road tunnel construction using neural networks. J. Constr. Eng. Manag. 138 (6), 679–687.

PMI, A., 2008. December). Guide to the Project Management Body of Knowledge: PMBOK 2000. Project Management Institute.

Polit, D.F., Beck, C.T., 2012. Nursing Research: Generating and Assessing Evidence for Nursing Practice, ninth ed. Wolters Klower Health, Lippincott Williams & Wilkins, Philadelphia, USA.

Radwan, H.G., 2013. Sensitivity analysis of head loss equations on the design of improved irrigation on-farm system in Egypt. Int. J. Adv. Res. Technol. 2 (1).

Ranasinghe, M., 2000. Impact of correlation and induced correlation on the estimation of project cost of buildings. Constr. Manag. Econ. 18 (4), 395–406.

Ratner, B., 2010. Variable selection methods in regression: ignorable problem, outing notable solution. J. Target. Meas. Anal. Mark. 18 (1), 65–75.

Rockwell, R.C., 1975. Assessment of multicollinearity: the Haitovsky test of the determinant. Socio. Methods Res. 3 (4), 308–320.

Runker, T.A., 1997. Selection of appropriate deffuzification methods using application specific properties. IEEE Trans. Fuzzy Syst. 5 (1), 72–79.

Shaheen Ahmed, A., Fayek, Aminah Robinson, Aminah Robinson, S.M., 2007. Fuzzy numbers in cost range estimating. J. Constr. Eng. Manag. 133 (4).

Shreenaath, A., Arunmozhi, S., Sivagamasundari, R., March 2015. Prediction of construction cost overrun in Tamil nadu- a statistical fuzzy approach. Int. J. Eng. Tech. Res. (IJETR) 3 (3), ISSN: 2321-0869.

Siddique, N., Adeli, H., 2013. Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing. John Wiley & Sons, Chichester, West Sussex.

Stevens, J.P., 2002. Applied Multivariate Statistics for the Social Sciences, fourth ed. Erlbaum, Hillsdale, NJ.

Stoy, C., Pollalis, S., Schalcher, H.-R., 2008. Drivers for cost estimating in early design: case study of residential construction. J. Constr. Eng. Manage. 134 (1), 32–39.

Stoy, C., Pollalis, S., Dursun, O., 2012. A concept for developing construction element cost models for German residential building projects. IJPOM Int. J. Proj. Organisat. Manag. 4 (1), 38.

Tabachnick, B.G., Fidell, L.S., 2007. Using Multivariate Statistics, fifth ed. Allyn & Bacon, US, Boston.

Wang, J., Ashuri, B., 2017 Jan 2. Predicting ENR construction cost index using machine-learning algorithms. Int. J. Constr. Educ. Res. 13 (1), 47–63.

Wang, Y.-R., Yu, C.-Y., Chan, H.-H., 2012. Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models. Int. J. Proj. Manag. 30 (4), 470–478.

Wilkinson, L., Dallal, G.E., 1981. Tests of significance in forward selection regression with an F-to enter stopping rule. Technometrics 23, 377–380.

Williams, T.P., Gong, J., 2014. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. Autom. ConStruct. 43, 23–29.

Woldesenbet, A., Jeong, H.S., 2012. Historical data driven and component based prediction models for predicting preliminary engineering costs of roadway projects. In: Proceedings of ASCE Construction Research Congress, pp. 417–426.

Yang, I.-T., 2005. Simulation-based estimation for correlated cost elements. Int. J. Proj. Manag. 23 (4), 275–282.

Zadeh, L.A., 1965. Fuzzy sets. Inf. Control 8 (3), 338–353.

Zadeh, L.A., 1973. Outline of a new approach to the analysis of complex systems and decision process. IEEE Trans. Syst. Man Cybern. 3, 28–44.

Zadeh, L.A., 1976. The concept of linguistic variable and its application to approximate reasoning – III. Inf. Sci. 9, 43–80.

Zhai, K., Jiang, N., Pedrycz, W., 2012. Cost prediction method based on an improved fuzzy model. Int. J. Adv. Manuf. Technol. 65 (5-8), 1045–1053.

Zhang, R., Ashuri, B., Shyr, Y., Deng, Y., 2018. Forecasting Construction Cost Index based on visibility graph: a network approach. Phys. Stat. Mech. Appl. 493, 239–252.