JOURNAL OF CIVIL ENGINEERING AND MANAGEMENT
ISSN 1392-3730/eISSN 1822-3605
2017 Volume 23(1): 85–95
https://doi.org/10.3846/13923730.2016.1210214

GENE EXPRESSION PROGRAMMING APPROACH TO COST ESTIMATION FORMULATION FOR UTILITY PROJECTS

Neda SHAHRARA[a], Tahir ÇELIK[b], Amir H. GANDOMI[c]

*[a]Department of Civil Engineering , Eastern Mediterranean University,
Famagusta North Cyprus via Mersin 10, Turkey
[b]Department of Civil Engineering, Cyprus International University,
Nicosia North Cyprus via Mersin 10, Turkey
[c]BEACON Center for the Study of Evolution in Action, Michigan State University,
East Lansing, MI 48824, USA*

Received 24 May 2015; accepted 25 Nov 2015

Abstract. This article utilizes gene expression programming (GEP) technique to develop a prediction model in order to automate estimating the construction cost of water and sewer replacement/rehabilitation projects. A database gathered for developing the model was established on the basis of data related to 210 actual water and sewer projects obtained from the City of San Diego, California, USA. To verify the predictability of the GEP model, it was examined to estimate the cost of the projects that were not included in the modelling process. Sensitivity analysis technique and professional experiences were employed to determine the contributions of the qualitative factors and quantifiable parameters affecting the cost estimate. The proposed model with correlation coefficient of 0.8467 is adequately capable of estimating the cost of water and sewer replacement/rehabilitation projects. The GEP-based design equation can easily be used for predesign purposes to help allocate budgets and available limited resources effectively.

Keywords: cost estimate, genetic programming, utility projects, water and sewer replacement/rehabilitation projects.

Introduction

Cost estimation is fundamental at feasibility study of infrastructure projects. Accurate estimation will help decision makers consider best alternatives without misconstruing technical and economic approaches. At the conceptual phase of a project the urgency of undertaking the project is explored, technical and funding options are evaluated, and objectives of the project are set (Wideman 1995).

In common form of an infrastructure project development, a public agency (owner) designs a project and invites the private sector firms (Contractors) to bid the construction of the project. Contract to undertake the project is awarded to the lowest bidder (DeCorla-Souza, Mayer 2010).

Cost estimate at the conceptual phase becomes cost and budget control baseline for both the owner and the Contractor (Hendrickson, Au 1998).

Reliable cost prediction, based on limited information at early stages of the planning phase of modernizing and upgrading infrastructure projects, becomes of grave importance to utilize limited available resources accordingly and allocate adequate budgets for their successful completion. Shehab *et al.* (2010) reported that according to the municipality officials' experience, in the past, unrealistically high cost estimates withheld the projects for future fiscal years, or low cost estimates resulted in inadequate budget allocation and constructing projects below ideal standards (Shehab, Farooq 2013). Recurrent incidents of sewer overflows into rivers and streams as well as water main breaks in the cities in the United States manifest that there is a dire need for upgrading aging and deteriorating drinking water and wastewater infrastructures. In fiscal year 2012, the U.S. Environmental Protection Agency (EPA) funded the Clean Water program $1.5 billion and the Drinking Water program $918 million from congressional appropriations. EPA grants capitalization funds to states of USA, which in turn provides low- or no-interest loans to local communities or utilities to pay for water distribution pipelines, treatment plants, sewer lines, and other similar infrastructure. EPA estimated funding requirements of almost $335 billion for drinking water infrastructure and $298 billion for wastewater infrastructure (Gómez 2013).

Various methodologies of machine learning techniques such as regression models and artificial intel-

Corresponding author: Amir H. Gandomi
E-mail: *a.h.gandomi@gmail.com*

Taylor & Francis
Taylor & Francis Group

ligence techniques can be employed for modelling a nonlinear system such as cost estimation. Two most renowned artificial intelligence methods used in nonlinear modelling are artificial neural networks (ANNs) (Haykin 1999) and Genetic Programming (GP) (Koza 1992; Gandomi *et al.* 2013).

Artificial intelligence methods have been widely used as prediction tools in recent decades. Review of comparative studies on artificial intelligence and traditional statistical techniques in various fields of applications shows that artificial neural networks outperform regression models as a tool for classification and prediction problems (Paliwal, Kumar 2009; Kim *et al.* 2004).

ANNs are one of the most well-known pattern recognition systems that are capable of learning from experience. ANNs are vastly used in cost estimating of building and infrastructure projects (Tatari, Kucukvar 2011). Several researchers attempted to develop cost estimation models in the earlier stages of developing infrastructure projects using regression models or ANNs. Hegazy and Amr (1998) used a neural network approach to develop a parametric cost estimating model for highway projects. Adeli and Wu (1998) formulated a regulation neural network based on a solid mathematical foundation for estimation of highway construction costs. Sodikov (2005) used ANN to analyse the impact of a different set of variables on the highway project cost and proposed a cost estimation technique for developing countries.

Successful usage of ANNs and regression models on cost estimation of aforementioned infrastructure projects encouraged some researchers to apply such models on cost estimation of sewer and water replacement or rehabilitation projects.

Using regression techniques, Clark *et al.* (2002) proposed seven separate cost estimating equations for water supply distribution models, summation of which would yield to the direct cost of replacing a new water distribution system. Besides shortfalls of regression techniques in comparison with other techniques, and the tedious procedure of using several models, indirect costs such as Contractor's overhead, profit, bonds, insurance and social costs were not taken into account.

Shehab *et al.* (2010) developed two models for utility rehabilitation projects using ANN and regression analysis and argued that ANN provided more accurate results.

Alex *et al.* (2010) developed cost prediction model using ANN for installation of water and sewer systems incorporating factors such as geographical location of the project, seasonal variation, average monthly temperature and historical construction cost data divided into four categories of labour, equipment, material and other costs. However, estimating the cost of mentioned four categories requires undertaking a detailed resource and productivity analysis as well as punctilious construction technology assessments which at the early stages of the studying the project seems to be abstract and superfluous.

Shehab *et al.* (2010) utilized ANN to develop a cost prediction model for installation of water and sewer systems using 50 historical data sets to evaluate the impact of six categories of pipes, sidewalks, manholes, pavement, soil, services and assemblies on the cost of the projects. Developing a model based on fewer sample projects does not yield a plausible and reliable model. Furthermore, despite promising application of ANNs on engineering problems, the process of obtaining a solution from available information is unknown and extracting practical prediction equations are not usually possible. Moreover, a neural network structure requires the researcher to predefine it (Alavi, Gandomi 2011).

Genetic algorithm (GA) is a robust optimization method based on the basic idea of genetics and natural selection. GA is considered to be efficiently applicable to vast spectrum of different engineering problems (Milani, G., Milani, F. 2008).

Genetic programming (GP) (Koza 1992) is a derivative of GA. GP solutions are computer programs in lieu of binary strings (Banzhaf *et al.* 1998). GP is a nonlinear structured alternative to fixed length solutions (Ferreira 2006). GP is based on Darwin's theory of evolution, expressed as "survival of the fittest". A group (population) of computer programs (individuals) continues reproducing with each other till the best individuals will survive and finally evolve to perform well in the specified scenario (Walker 2001). There are wide-range applications of GP in prediction, optimization and classification problems in both science and engineering domains (Yaghouby *et al.* 2010, 2012; Gandomi, Roke 2014).

GP's ability to develop simple prediction equations with no need to considering an existing relationship is its main superiority over the conventional statistical and ANN techniques (Gandomi *et al.* 2012). GP is intrinsically capable of finding the best solution by evaluating fitness of the computer programs over numerous generations; on the contrary, as mentioned earlier, in ANN, data should be normalized at the outset, and best network architecture first should be established (Gandomi, Roke 2015).

When the analyst creates an equation, applicability and validity of the cost estimation model is more discernible since an equation can check with common sense especially in the case of proposals requiring acquisition of management and owner approval (Smith, Mason 2010).

Gene expression programming (GEP) is a recent variant of GP. The GEP is able to evolve computer programs of different sizes and shapes. GEP is extremely adaptable and supersedes the existing evolutionary techniques (Ferreira 2001). Several scientists applied GEP to civil engineering realm (Azamathulla 2013; Alavi, Gandomi 2011; Gandomi *et al.* 2011; Azamathulla, Ahmad 2013).

This study utilized the GEP technique to build a predictive model for cost estimation of water and sewer utility rehabilitation and replacement infrastructure projects to our best knowledge for the first time. The developed model considers readily available variables with substantial impact on the cost of the projects. Sensitivity analysis technique and professional experiences were employed to determine the contributions of the qualitative factors and quantifiable parameters affecting the cost estimate.

1. Genetic programming

GP, an extension of GA, was invented by Cramer (1985) and further developed by Koza (1992) and Ferreira (2006). Although GP applies most of key ideas of GA and uses GA operators such as selection, crossover and mutation with slight modifications, its nonlinear structure creates a more versatile system of representation than that of GA (Ferreira 2006; Gandomi *et al.* 2012). GP produces computer programs with dynamic variability and hierarchical character presented in form of parse trees (Koza 1992). A population member in hierarchically structured tree-based GP composes of functions and terminals selected from a set of functions and a set of terminals. Figure 1 is an illustration of a simple tree-based GP model (Gandomi *et al.* 2012). GP can be implemented using any programming language (like LISP) capable of working with computer programs as data and linking, compiling and executing new programs (Koza 1992). The GP represents basic structures of the approximation model along with values of its parameters; however GA solutions are fixed length strings of numbers. The GA, similar to other traditional optimization techniques, is used in parameter optimizations to evolve the best values for a given set of model parameters (Javadi, Rezania 2009; Alavi, Gandomi 2011).

The GP optimizes a population of computer programs in terms of a fitness landscape which defines how good a candidate solution (program) to achieve the set aim is; in other words, GP intends to optimize the fitness function, a particular objective function, which is used to evaluate the fitness of each program (Alavi, Gandomi 2011).

GEP is a linear extension of GP comprised of autonomous entities of genotype and phenotype. In genetics, an organism's complete hereditary information is called genotype; and an organism's actual observed properties, such as morphology, development, or behaviour is called phenotype. Ferriera (2001) translated the language of chromosomes into the language of expression tree (ET), a tree-like structure.

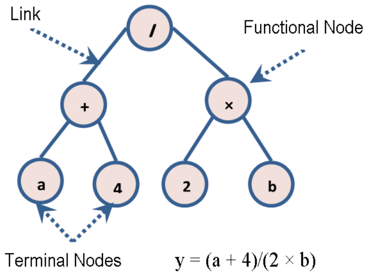

Fig. 1. Illustration of a GP model in a tree-like structure

1.1. Gene expression programming (GEP)

The GEP is a natural development of GP first invented by Ferreira (2001). The GEP consists of five main components: 1) function set; 2) terminal set; 3) fitness function; 4) control parameters; and 5) termination condition (Gandomi *et al.* 2012).

In GEP, individuals are linear strings of fixed length (the genome or chromosomes) which later are represented in form of nonlinear structures of different sizes and shapes (phenome, i.e. expression trees (ETs)). Since genotype and phenotype of an individual are independent, only the genome is carried to the next generation. Respectively, replication and mutation of the structures are not required any more (Ferreira 2006).

Therefore, the main players in GEP are the chromosomes and ETs. An advantage of the GEP technique is that the creation of genetic diversity is extremely simplified because genetic operators work at the chromosome level (Gandomi *et al.* 2012).

Furthermore, multigenic nature of GEP forms complex multisubunit expression trees (ETs) (programs) which are both separate entities and part of a more complex, hierarchical structure at the same time (Ferreira 2001). Each GEP gene contains a list of symbols with a fixed-length that can be any element from a function set like {þ, −, ×,⁄, Log} and the terminal set like {a, b, c, 3} (Gandomi *et al.* 2012).

Ferreira (2001) created a new language of GEP, called Karva language to read and express the information encoded in the chromosomes which, as an important feature of GEP, are capable of representing any pars-tree.

The mathematical expression below:

$$(4 \times a) / (2 + \cos(b + c)), \tag{1}$$

can be expressed in Karva language as follows:

$$/\times + 4a2\cos + bc, \tag{2}$$

where *a*, *b*, and *c* – variables; and 2 and 4 – constants. The variables or constants used in a problem are called terminals. This GEP gene can be illustrated as an ET shown in Figure 2. This kind of expression is the phenotype of GEP individuals (Ferreira 2001).

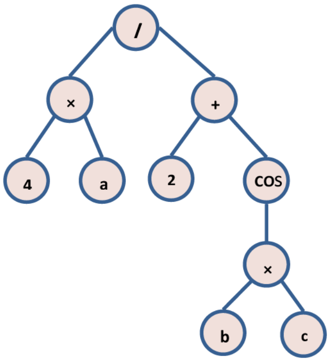

Fig. 2. Typical illustration of an ET

The conversion starts from the first position in the K-expression, which corresponds to the root of ET, and reads through the string one by one.

By recording the nodes from left to right in each layer of the ET, from root layer down to the deepest one, an ET can be expressed in K-expression. The assemblage of ET is complete when the deepest layer is composed only of terminals, meaning that there is no longer any function left to make a link to any terminal. The transfer of information from a gene into an ET is called translation (Ferreira 2001). Since GEP chromosomes comprise of predetermined fixed length genes, the only variable would be the size of the corresponding ETs, meaning that some elements are not useful for the genome mapping. Therefore, the acceptable length of a K-expression may be equal or less than the length of the GEP gene. GEP applies a head-tail method to secure the validity of a randomly selected genome. Each GEP gene is composed of a head and a tail. The head may contain both function and terminal symbols, whereas the tail may contain terminal symbols only (Alavi, Gandomi 2011; Ferreira 2001).

Random generation of each individual's fixed-length chromosomes creates the initial population of GEP model. Later on, the chromosomes are expressed, and the fitness of each individual is examined. The individuals with better fitness are then selected to reproduce with modification. The individuals of this new generation are subjected to the same developmental process: expression of the genomes, confrontation of the selection environment, and reproduction with modification. The previous process is repeated for a definite number of generations until a solution has been found. By roulette wheel sampling (with elitism) method, the individuals with better fitness are selected and replicated into the next generation. This secures the survival and cloning of the best individual to the next generation (Alavi, Gandomi 2011). Figure 3 demonstrates the basic steps of GEP (Ferreira 2001).

In this study, the GEP approach was utilized to acquire a valid relationship between the cost of sewer and water replacement/rehabilitation projects and impacting variables.

2. Data preparation

This study is proposing a predictive model for cost estimation of rehabilitation and/or replacement of sewer and water projects utilizing GP technique leading to improved results as well as simplified procedures. To develop the prediction model, 210 actual proposals related to water and sewer projects submitted by the lowest bidders to the City of San Diego, CA, USA (1999–2013) were obtained. Basically, The City of San Diego designs the utility systems. The design may be performed in house; or outsourced by hiring a private engineering company. Afterwards the City invites prequalified construction Contractors to bid the designed project. The bid which is submitted by the competing Contractors is based on bill of quantities; and comprises of itemized component values, The City awards each construction contract to the lowest, qualified bidder. Since each project's cost estimation includes many components such as pipes, manholes, pavement, curb and gutter, water pollution control plan, etc., the most important items with higher impact on the outcome, which could be easily assessed in the conceptual stage also, were chosen as inputs to the GP model. These items were selected by the aid of sensitivity analysis and expert judgment. Sensitivity analysis provides a useful tool for analysing the impact of changes in input variables in terms of bid evaluation (Grimsey, Lewis 2004). Sensitivity analysis carries out a clear and adjustable procedure by varying the parameters randomly one at a time to observe the impact of changes on the outcome. For any given parameter a number of incremental changes are made and the final indicator value (outcome) is computed each time recording the degree of change from its baseline (Jenkins *et al.* 2011). Variables, which their variation could have a substantial impact on the projects' outcome, will be separated as alternative input variables for GP model. Afterwards professional judgment/experience is utilized to choose among the variables that can be readily and accurately assessed at the conceptual stage. There are a few qualitative factors that can impact productivity such as soil classification, pavement condition, traffic and finally seasonal effect. The latter's impact is not substantial in San Diego since no dramatic weather

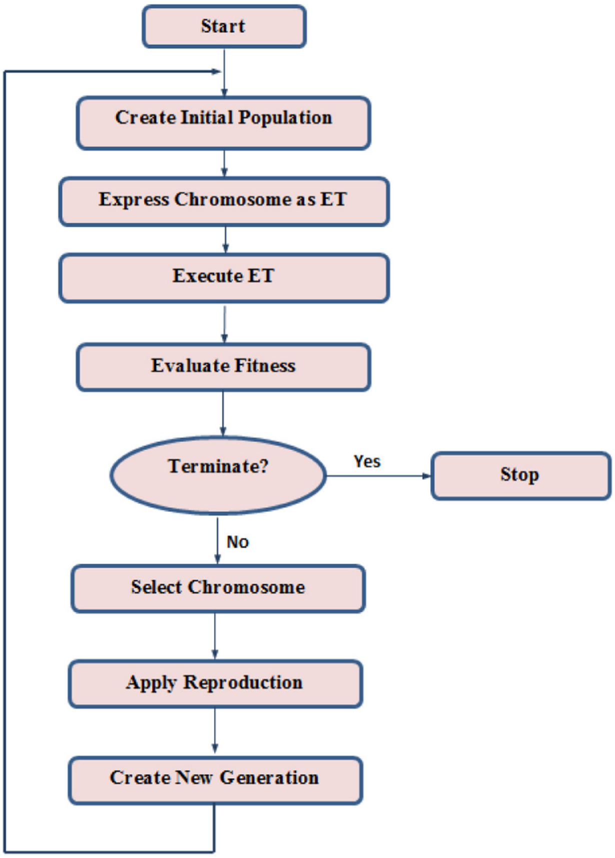

Fig. 3. Basic representation of the GEP algorithm

fluctuation is observed in San Diego's weather forecast. Qualitative factors mentioned above were identified by evaluating corresponding project bid items to incorporate in developing GP cost estimation model.

2.1. Data analysis

While most of the projects studied in this research involved the replacement/rehabilitation of both sewer and water mains in a neighbourhood, some jobs were exclusively water or sewer replacements or sewer rehabilitation. Prevalently utilized construction method was excavation (open trench) replacement; and cured-in-place pipe (CIPP) rehabilitation method. Occasionally different rehabilitation technologies such as close fit lining, and slip lining or trenchless replacement methods such as boring and pipe bursting were applied.

Trenchless methods are more expensive in comparison to open trench methods; they are the best option for the installation of pipelines under a road, railroad, freeway, or in other situations where trenching is not possible; there is little business and human costs (social costs) associated with traffic congestion, restriction of access, dirt, noise, air pollution. Natural habitats and landscaping will remain undisturbed. Therefore, revegetation and erosion control provisions will not be required. These methods are less labour intensive with faster completion (EPA 1999). But it is observed that in the City of San Diego, these methods have been utilized only where there was not a possibility of implementing an open trench technology therefore the common practice was assumed to be open trench and CIPP if it was possible, thus the construction technology selection was not an influential factor in model development.

To perform sensitivity analysis using what–if analysis function on an Excel spread sheet following actions were taken:

- Bringing all data pertinent to each project in one spread sheet since the City of San Diego announces information regarding each bid result such as bid items, quantities, unit costs and proposed cost by the lower bidder in separate spread sheets.
- Identifying the variables (bid items) which seem to have a significant impact on the outcome of the projects (and can be easily assessed in the conceptual stage).
- Identifying a likely range for these variables, centered on the most likely assumed values.
- Calculating the impact of different combinations of these variables on the total cost of the projects (Rogers, Duffy 2012).

Normally each proposed project is broken down into approximately 110–140 bid items. By aid of sensitivity analysis, 24 bid items were identified to be most influential on the cost estimation of the projects. The 24 items were grouped into 4 categories consisting of:

1. Replacement/rehabilitation of sewer and water mains.
2. Installation of manholes, sewer laterals, private replumbs and water services with various diameters and thicknesses.
3. Pavement conditions including asphalt concrete, concrete pavement replacement, temporary resurfacing, slurry seal, asphalt concrete patching, pavement removal, crack sealing, pavement fabric, cold milling, pavement restoration adjacent to trench, striping, extra thick pavement removal. According to professional experience, most of the field order allowance (which usually is listed as a bid item) ends up to be allocated for pavement repair purposes because most of the time, the pavement condition is not objectively evaluated on the conceptual stage; therefore field order item was taken into account under pavement category.
4. Soil conditions taking into account the soil type impact and proposed costs of shoring, dewatering and pipe installation since if the soil condition declines, the installation would be more difficult, slower, labour incentive and cost of activities such as shoring and dewatering would increase and consequently the price allocated for overall installation would rise.
5. Traffic control including traffic control plans and set up cost and studies conducted on the neighbourhood's traffic conditions.

The formulas below are proposed to be used to simplify the input variables for GP model where:

$S_{\text{diameter (inches), pipe type}}$ = Sewer Main Length (Linear Feet); $S_{\text{diameter (inches) r}}$ = CIPP Sewer Main Length (Linear Feet); $W_{\text{diameter (inches)}}$ = Water Main Length (Linear Feet); $SL_{\text{diameter (inches)}}$ = Number of Sewer Laterals; $WS_{\text{ diameter (inches)}}$ = Number of Water Services, MH = Number of Manholes.

Input Variables to GP model are listed below:

X_1: Soil Condition (1 = best through 10 = least desirability according to the table of relative desirability of soils (MultiQuip Inc 2011));

X_2: Pavement Condition (1: Good (allocated cost per category (3) items lower than 10% of total cost) 2: Average (between 10–25%) 3: Bad (above 25%));

X_3: Traffic Control (1: moderate, 2: busy).

Each bid item's quantity related to a certain pipe (with different size and property) can be used as one input variable for GP model; but this way the number of input variables will unnecessarily be numerous. In order to simplify GP model, Eqn (3) is proposed to combine quantities of different pipes with different sizes and properties to bring certain bid items (that have linear relation with each other) together and generate one input for GP model in lieu of numerous input variables, where the parameter values in Eqn (3) are the average unit price of related item per unit price of 8" sewer main item during 1999–2013. And Subscripts "s" and "c" stand for Schedule and Class PVC pipes respectively.

$$X_4 = 2.65S_{27} + 2.04\left(S_{24} + S_{18}\right) + 1.74S_{15} + 1.63S_{12} + 1.37S_{10} + S_8 + 1.45S_{8s} + 1.34W_{16} + 1.26W_{12c235} + 1.48W_{12c150} + W_8 + 0.74 \left(S_{61} + 1.25S_{8r} + 1.5S_{10r} + 1.75S_{12r} + 10S_{36r}\right). \tag{3}$$

Similar to Eqn (3), in order to reduce the number of input variables to one input, Eqn (4) was used to combine the number of sewer laterals, water services, and manholes with different sizes and properties to generate one input for GP model, where the parameter values in Eqn (4) are the average unit price of related item per unit price of 4” sewer lateral during 1999–2013. Subscripts “pr”, “n”, and “r” stand for private re-plumb, new and rehab respectively:

$$X_5 = 1.3SL_6 + SL_4 + 5.6SL_{pr} + 2.09WS_2 + 0.97WS_1 + 5MH_n + 2.74MH_r. \tag{4}$$

Nominal values of the projects’ proposed costs by the lowest bidders were converted to real values by the relevant price indexes released by U.S. Department of Labour, Bureau of Labour Statistics (Coinnews Media Group LLC 2014). It is worth mentioning that the outcome of GP formula would yield a cost estimate in real prices; in order for the user to come up with the nominal cost estimate, the outcome should be brought back to nominal values.

2.2. Database

The model was developed based on 210 sets of data related to sewer and water replacement/ rehabilitation projects obtained from the City of San Diego, CA, USA. The essential objective of a Machine Learning approach is to find solutions that perform well not only on the cases used for learning but also on cases of new unseen data. This is known as generalization ability, and failure to fulfil this is called overfitting (Goncalves, Silva 2011). Overfitting is usually the result of excessively trained algorithm which in spite of decreasing the training error, it increases the testing error rapidly (Gandomi *et al.* 2012). An efficient approach to prevent overfitting and improve generalization of the model is to test the derived models on a validation set to achieve a better generalization (Banzhaf *et al.* 1998) which was employed in this study. Correspondingly, the available data sets were randomly divided into learning, validation, and testing subsets. To perform genetic evolution, the learning data were used for training purposes. To determine the generalization capability of the models on the untrained data, the validation data were used for model selection purposes. Training data alluded to learning and validation data which both were involved in the modelling process. Finally, as the outcome of the runs, the model with best performance on both of the learning and validation data sets is selected. To examine performance of the optimal model derived from GP on unseen data, the testing data were engaged which had no affiliation with building the models.

In order to achieve a uniform data division, several combinations of the training and testing sets were selected in a way that the statistical properties of the involved parameters (e.g., maximum, minimum, and mean) were consistent in the training and testing data sets (Gandomi *et al.* 2012). Out of the 210 data sets, 185 data vectors were taken for the training process (160 sets for learning and 25 sets for validation). The remaining 25 sets were used for the testing of the derived model.

3. Model development

3.1. Performance measures

Selection of the best model was based on the strategies below (Gandomi *et al.* 2012):

1. The simplest model, although this was not a main factor, which was controlled by the user through the parameter settings (e.g., number of genes or head size);
2. The model with the best fitness value on the learning data;

and

3. The model with the best fitness value on the validation data.

The best GP model was inferred by minimizing the following objective function (OBJ) which was used to verify acceptability of predicted output versus the actual bid proposals.

$$\mathrm{OBJ} = \left(\frac{\mathrm{No.}_{\mathrm{Learning\text{-}No.}_{\mathrm{Valitating}}}}{\mathrm{No.}_{\mathrm{Training}}}\right)\rho_{\mathrm{Learning}} + \frac{2\mathrm{No.}_{\mathrm{Validation}}}{\mathrm{No.}_{\mathrm{Training}}}\rho_{\mathrm{Validation}}, \tag{5}$$

where No.Training, No.Learning, and No.Validation are respectively, the number of training, learning, and validation data and ρ is the performance index as follows (Gandomi, Roke 2013):

$$\rho = \frac{\mathrm{RRMSE}}{1+R}. \tag{6}$$

The RRMSE and R are widely used parameters for the performance measurement respectively, the root mean squared error, mean absolute error, and correlation coefficient (Milani, Benasciutti 2010). The following equations were used to determine the RRMSE and R values:

$$\mathrm{RRMSE} = \frac{1}{\left|\overline{h_i}\right|}\sqrt{\frac{\sum_{i=1}^{n}\left(h_i^2 - t_i^2\right)}{n}}; \tag{7}$$

$$R = \frac{\sum_{i=1}^{n}(h_i - \overline{h_i})\left(t_i - \overline{t_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(h_i - \overline{h_i}\right)^2 \sum_{i=1}^{n}\left(t_i - \overline{t_i}\right)^2}}, \tag{8}$$

where h_i and t_i are respectively, the actual and calculated outputs for the i^{th} output, $\overline{h_i}$ and $\overline{t_i}$ are average of the actual and calculated outputs; and n – number of samples. Since R value does not change by equal shifting of the output values predicted by a model, it is acknowledged not to be a good indicator on its own for evaluating the accuracy of a model on its own. On the other hand, besides assuming the impact of various data divisions for the learning and validation data, the performance index (ρ) simultaneously takes into account the changes of RRMSE and R. Lower RRMSE and higher R values yield in lower OBJ indicating a more accurate model (Gandomi, Roke 2015). The values obtained for R, RRMSE, and ρ are respectively, 0.8467, 0.4065, and 0.220.

3.2. Model development using GP

Several preliminary runs were made to observe the performance. The number of programs in the population that GP evolves is set by the population size (number of chromosomes). A run takes longer with a larger population size. The proper number of population depends on the number of possible solutions and complexity of the problem (Gandomi *et al.* 2012).

Three optimal levels were set for the population size (50, 150, and 300). The architecture of the models evolved by GP is determined by head size and number of genes. The head size determines the complexity of each term in the evolved model. The number of terms in the model is determined by the number of genes per chromosome. Each gene codes for a different sub-ET. Five optimal levels were considered for the head size and number of genes. For the number of genes greater than one, the addition and multiplication linking functions were used to link the mathematical terms encoded in each gene. There are 2 (linking functions) × 4 (head size) × 5 (number of genes) = 40 different combinations of the parameters. All of these combinations were tested, and replications for each combination were carried out. Table 1 demonstrates parameter settings for GP algorithm.

Table 1. Parameter settings for the GP algorithm

Parameter	Settings
General	
Chromosome	50,150, 300
Genes	2, 4, 6, 10, 12
Head size	2–6
Tail size	9
DC size	9
Gene size	23
Linking function	∑, ∏
Genetic operator	
Mutation rate	0.044
Inversion rate	0.1
IS transposition rate	0.1
RIS transposition rate	0.1
One-point recombination rate	0.3
Two-point recombination rate	0.3
Gene recombination rate	0.1
Gene transposition rate	0.1
Numerical Constants	
Constants per gene	9
Data type	Flouting-point
Lower bound	–10
Upper bound	10

The period of time acceptable for evolution to occur without improvement in best fitness is set through the generations without change parameter. In this study, basic arithmetic operators and mathematical functions were utilized to get the optimum GP model. The mean absolute error function was used to calculate the overall fitness of the evolved programs. On this GP model variable pressure function (variable pressure = 0.01) has been also employed.

The program was run until there was no longer significant improvement in the performance of the models. The GP algorithm was implemented using GeneXproTools (2014).

4. Results and discussion

4.1. GP-Based formulation

The GP-based formulation of project cost estimation (K\$) in terms of x_1, x_2, x_3, x_4, x_5 is as follows:

$$\begin{aligned}\text{Cost} = {} & x_4\left(57.5303x_2 + 151.7352\right) + \\ & \left(8.9097x_2 - 42.7632\right)^2 - 0.3472 \\ & \left(x_4 + 1.4305\right)\left(x_4 - x_5\right) + \frac{x_1.x_3^3}{\left(x_4 - 0.386\right)}.\end{aligned} \tag{9}$$

The formulation mentioned above displays a complex arrangement of operators, variables, and constants that are used to predict cost estimation. The expression tree of the derived equation is given in Figure 4. The proposed equation is composed of four independent subprograms (genes) interrelated by the addition operator. Embodying a particular character of the problem, each subprogram adds a distinct function to the developed solution (Ferreira 2001). In other words, each evolved subprogram contains important information about the physiology of the final model (Gandomi *et al.* 2012). Each gene, as a part of the final equation, is engaged to address a certain aspect of the problem.

4.2. Comparison of GP model with the City of San Diego's engineering estimate

The City of San Diego announces a suggested cost estimate for each project in the project's bid documents. A project cost estimate, called engineering cost estimate, is attained through an in-house lengthy and rather expensive system which operates on an educated guess based on the past bids and the judgement of the project manager who puts the project bid documents together. Most of the time the engineering estimate is so much higher than the lowest bidder's proposed price or is lower. As mentioned before, engineering estimate becomes a gauge for budget allocation. Unrealistically high cost estimate prevents the project from being implemented or low budget allocation results in many shortfalls in the future. Besides, the

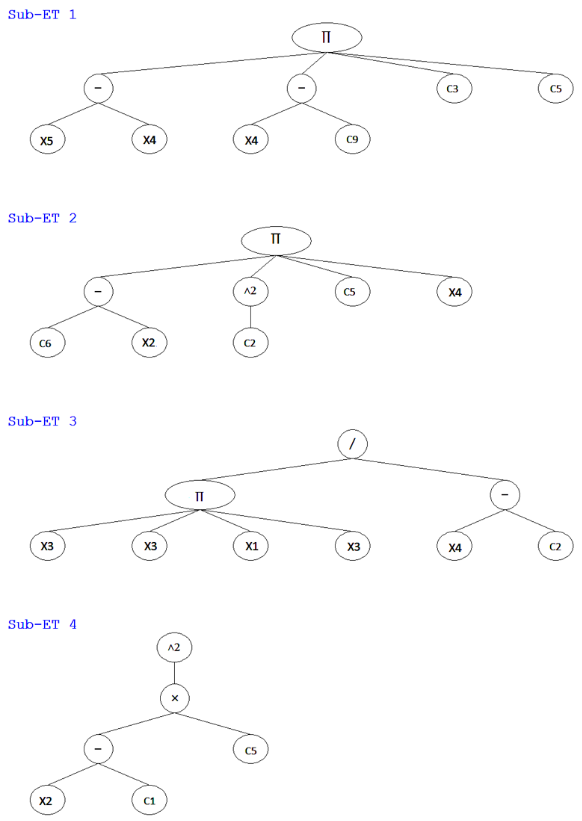

Fig. 4. Expression tree for cost estimation of sewer and water projects

announced engineering estimate somehow gives direction to the bidders. When an unrealistically high engineering estimate is announced, the Contractors are inclined to inflate their bid price to earn a bigger profit margin.

To confirm that the proposed GP model would be a capable tool, the results driven from GP formulation were compared to the City's engineering estimates. Comparison of the results of GP equation with the engineering estimate proves the outperformance of the GP equation. Besides an improved accuracy of GP equation, its usage is very easy. The formula is built on the basis of extensive sets of data with plenty of possible real life scenarios. Furthermore, GP model takes into account qualitative factors that could impact productivity such as soil classification, pavement condition, and traffic.

4.3. Model validity

According to Smith (1986), there is a solid correlation between the predicted actual values if a model maintains $R > 0.8$. Should the MAE values be at the minimum, the solution is considered reliable (Gandomi *et al.* 2011).

The results demonstrates that the proposed GP model with low RMSE and MAE and high R values is able to predict the target values with satisfactory accuracy. Reliable predictive ability and generalization performance of the model is concluded from the good performance of the model on the training (learning and validation) and testing data.

Table 2. Statistical parameters of the GP model for the external validation

Item	Formula	Condition	GP
1	R	$0.8 < R$	0.8467
2	$k = \left[\sum_{i=1}^{n}(h_i \times t_i)\right] / h_i^2$	$0.85 < k < 1.15$	1.0008
3	$k' = \left[\sum_{i=1}^{n}(h_i \times t_i)\right] / t_i^2$	$0.85 < k' < 1.15$	0.8950

Greatly affecting the accuracy of the final models, the amount of data used for the modeling process gains importance (Gandomi *et al.* 2012). Frank and Todeschini (1994) stated that a model can be considered acceptable if the minimum ratio of the number of objects per the number of selected variables is 3 preferably 5 yielding to more accurate solution. In this study, this ratio is as high as 160 ⁄ 5 = 32.

To examine external verification of the GP model on the testing data sets, Golbraikh and Tropsha's suggestion, that at least one slope of regression lines (k or k_0) through the origin should be close to 1, was checked as well (Golbraikh, Tropsha 2002).

The considered validation criteria and the pertinent outcomes acquired by the proposed model are presented in Table 2. Derived model satisfies the required conditions. The validation phase justifies soundness and strength of the prediction model.

The main feature of the proposed GP-based model is that it can readily be implemented by using the attainable accurate information with substantial impact on the project cost. Furthermore qualitative factors which affect productivity such as traffic, soil classification and pavement condition are incorporated into the model.

Most of the existing prediction models rely on assuming the structure of the model in advance, which may fall short. Thus, they cannot efficiently consider the interactions between the dependent and independent variables (Gandomi, Alavi 2011b).

On the other hand, GP produces clear relationships for project's cost estimation without assuming prior forms of the existing relationships. It directly learns from data presented to them. This is the same task followed by ANNs and other soft computing techniques (Gandomi *et al.* 2012).

A remarkable advantage of GP over ANNs is that it generates a transparent and structured representation of the system studied. Because of the large complexity of the network structure, ANNs do not give a transparent function relating the inputs to the corresponding outputs (Gandomi, Alavi 2011a).

4.4. Variable importance

The relative importance of each predictor variable in the GP analysis can be assessed on GP model. GeneXpro-

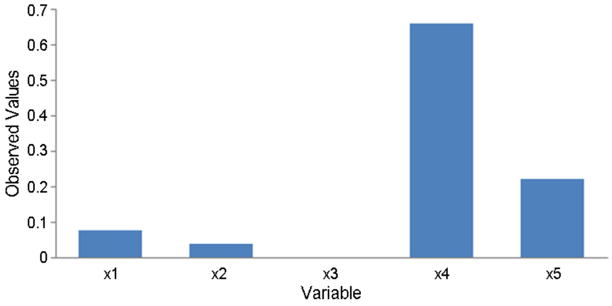

Fig. 5. Contributions of the predictor variables in the GP analysis

Tools computes the variable importance of all the variables in the model by randomizing its input values and then computing the decrease in the R-square between the model output and the target. The results for all variables are then normalized in order that they add up to 1 (GeneXproTools 2014). The variable importance of the predictor variables are displayed in Figure 5. As it is shown, the sewer and water mains incur the highest cost which is known from professional point of view too.

4.5. Parametric study

A parametric analysis was performed in this study to verify the robustness of GP-based prediction equation. The methodology is to change only one parameter at the time while other parameters are kept constant at the average values of their entire data sets. Figure 6 presents the parametric analysis of cost estimation in the GP model. An expected behaviour pattern is seen in the Figure 6. According to reported professional experience too sewer and water main installations incur most of the cost of a project.

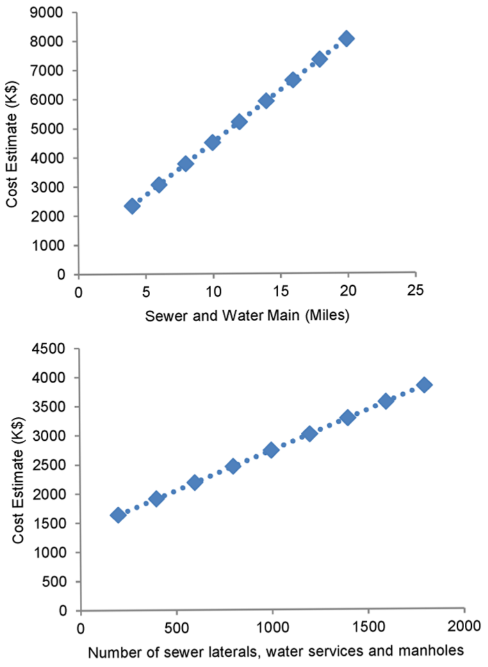

Fig. 6. Parametric analysis of cost estimate in the GP model

Conclusions

GEP, a variant of GP, was utilized to formulate the cost estimation of sewer and water rehabilitation/replacement projects. The proposed model, serving as a successful prediction tool, was developed based on data pertaining to 210 sewer and water replacement/rehabilitation projects from year 1999 to 2013 acquired from the City of San Diego, California, USA.

The conclusions below are drawn from this research:

1. Validity of the model was examined on testing data sets which were not part of training data sets. The GP prediction model efficiently satisfied the conditions of different criteria considered for its external validation as well.
2. The developed system offers an improved cost estimation model with higher accuracy in comparison with the owner's published engineering estimates; however our model is an explicit formula.
3. Unlike engineering estimates, using such a simple formula opts out the need to go through expensive and protracted cost estimation process on the conceptual stage of a project assessment
4. The GP cost estimation formula gives a simple solution with fairly less inputs which are easily attainable at the conceptual stage of a project assessment.
5. GP model takes into account the qualitative productivity factors such as traffic, soil and existing pavement conditions.
6. This model will lead to a more objective resource allocation for funding and decision making purposes and offers a more accurate cost baseline for both bidders and the City.

Acknowledgements

The first author would like to thank Jose Ortiz and Bijan Shakiba for sharing their invaluable experiences.

Disclosure statement

Authors do not have any competing financial, professional, or personal interests from other parties.

References

Adeli, H.; Wu, M. 1998. Regulation neural network for construction cost estimation, *Journal of Construction Engineering and Management* 124(1): 18–24. https://doi.org/10.1061/(ASCE)0733-9364(1998)124:1(18)

Alavi, A. H.; Gandomi, A. H. 2011. A robust data mining approach for formulation of geotechnical engineering systems, *Engineering Computations* 28(3): 242–274. https://doi.org/10.1108/02644401111118132

Alex, D. P.; Al Hussein, M.; Bouferguene, A.; Fernando, S. 2010. Artificial neural network model for cost estimation: city of Edmonton's water and sewer installation services, *Journal of Construction Engineering and Management* 136(7): 745–756. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000184

Azamathulla, H. 2013. Gene-expression programming to predict friction factor for southern Italian rivers, *Neural Computing and Applications* 23(5): 1421–1426. https://doi.org/10.1007/s00521-012-1091-2

Azamathulla, H.; Ahmad, Z. 2013. Estimation of critical velocity for slurry transport through pipeline using adaptive neuro-fuzzy interference system and gene-expression programming, *Journal of Pipeline Systems Engineering and Practice* 4(2): 131–137. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000123

Banzhaf, W.; Nordin, P.; Keller, R.; Francone, F. 1998. *Genetic programming – an introduction. On the automatic evolution of computer programs and its application.* San Francisco, California/Heidelberg, Germany: dpunkt/Morgan Kaufmann.

Clark, R. M.; Sivaganesan, M.; Selvakumar, A.; Sethi, V. 2002. Cost models for water supply distribution systems, *Journal of WaterRresoureces Planning and Management* 128(5): 312–321. https://doi.org/10.1061/(ASCE)0733-9496(2002)128:5(312)

Coinnews Media Group LLC. 2014. *US inflation calculator* [online], [cited 20 April 2015]. Available from Internet: http://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/

Cramer, N. L. 1985. *A representation for the adaptive generation of simple sequential programs.* s.l., Erlbaum.

DeCorla-Souza, P.; Mayer, J. 2010. *Public–Private Partnership Concessions for highway projects.* U.S. Department of Trasportation, Washington, DC.

EPA. 1999. *Collection systems O&M fact sheet: Trenchless sewer rehabilitation*. United States Environmental Protection Agency, Washington, D.C.

Ferreira, C. 2001. Gene expression programming: a new adaptive algorithm for solving problems, *Complex Systems* 13(2): 87–129.

Ferreira, C. 2006. Automatically defined functions in gene expression programming, *Genetic Systems Programming: Theory and Experiences, Studies in Computational Intelligence* 13: 21–56. https://doi.org/10.1007/3-540-32498-4_2

Frank, I. E.; Todeschini, R. 1994. *The data analysis handbook.* Amsterdam: Elsevier.

Gandomi, A. H.; Alavi, A. H. 2011a. Applications of computational intelligence in behavior simulation of concrete materials, *Computational Optimization & Applications*, Volume 359 of the series Studies in Computational Intelligence, 221–243. http://dx.doi.org/10.1007/978-3-642-20986-4_9

Gandomi, A. H.; Alavi, A. H. 2011b. Multi-stage genetic programming: a new strategy to nonlinear system modeling, *Information Sciences* 181(23): 5227–5239. https://doi.org/10.1016/j.ins.2011.07.026

Gandomi, A. H.; Alavi, A. H.; Mirzahosseini, M.; Nejad, F. 2011. Nonlinear genetic-based models for prediction of flow number of asphalt mixtures, *Journal of Materials in Civil Engineering* 23(3): 248–263. https://doi.org/10.1061/(ASCE)MT.1943-5533.0000154

Gandomi, A. H.; Babanajad, S.; Alavi, A. H.; Farnam, Y. 2012. Novel approach to strength modeling of concrete under triaxial compression, *Journal of Materials in Civil Engineering* 24(9): 1132–1143. https://doi.org/10.1061/(ASCE)MT.1943-5533.0000494

Gandomi, A. H.; Roke, D. 2013. Intelligent formulation of structural engineering systems, in *Seventh M.I.T. Conference on Computational Fluid and Solid Mechanics – Focus: Multiphysics & Multiscale*, Massachusetts Institute of Technology, Cambridge, MA.

Gandomi, A. H.; Roke, D. 2014. Seismic response prediction of self-centering concentrically braced frames using genetic programming, *Structures Congress* 2014: 1221–1232. https://doi.org/10.1061/9780784413357.110

Gandomi, A. H.; Roke, D. 2015. Assessment of artificial neural network and genetic programming as predictive tools, *Advances in Engineering Software* 88: 63–72. https://doi.org/10.1016/j.advengsoft.2015.05.007

Gandomi, A. H.; Yang, X.; Talatahari, S.; Alav, A. 2013. *Metaheuristic applications in structures and infrastructures.* 1st ed. Waltham, MA, Elsevier.

GeneXproTools. 2014. *Gepsoft* [online], [cited 20 April 2015]. Available from Internet: http://www.gepsoft.com/GeneXproTools/AnalysesAndComputations/VariableImportance.htm

Golbraikh, A.; Tropsha, A. 2002. Beware of q2!, *Journal of Molecular Graphics and Modelling* 20(4): 269–276. https://doi.org/10.1016/S1093-3263(01)00123-1

Gómez, J. A. 2013. *Water infrastructure approaches and issues for financing drinking water and wastewater infrastructure*. United States Government Accountability Office, Washington.

Goncalves, I.; Silva, S. 2011. Experiments on controlling overfitting in genetic programming, in *15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, 10–13 October 2011, Lisbon, Portugal, 152–156.

Grimsey, D.; Lewis, M. K. 2004. *Public private partnerships.* Northampton, MA: Edward Elgar Publishing, Inc. https://doi.org/10.4337/9781845423438

Haykin, S. 1999. *Neural networks-a comprehensive foundation.* 2nd ed. Englewood Cliffs, NJ: Prentice Hall.

Hegazy, T.; Amr, A. 1998. Neural network model for parametric cost estimation of highway projects, *Journal of Construction Engineering and Management* 124(3): 210–218. https://doi.org/10.1061/(ASCE)0733-9364(1998)124:3(210)

Hendrickson, C.; Au, T. 1998. *Project management for construction: fundamental concepts for owners, engineers, architects and builders.* USA: Prentice Hall.

Javadi, A. A.; Rezania, M. 2009. Applications of artificial intelligence and data mining techniques in soil modeling. *Geomechanics and Engineering: An International Journal* 1(1): 53–74. https://doi.org/10.12989/gae.2009.1.1.053

Jenkins, G.; Kuo, C. Y.; Harberger, A. C. 2011. *Cost-Benefit analysis for investment decisions.* Queen's University, Kingston.

Kim, G.; An, S.; Kang, K. 2004. Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning, *Building and Environment* 39(10): 1235–1242. https://doi.org/10.1016/j.buildenv.2004.02.013

Koza, J. R. 1992. *Genetic programming: on the programming of computers by means of natural selection.* Cambridge, MA: MIT Press.

Milani, G.; Benasciutti, D. 2010. Homogenized limit analysis of masonry structures with random input properties: polynomial responsesurface approximation and Monte Carlo simulations, *Structural Engineering and Mechanics* 34(4): 417–445. https://doi.org/10.12989/sem.2010.34.4.417

Milani, G.; Milani, F. 2008. Genetic algorithm for the optimization of rubber insulated high voltage power cables production lines, *Computers & Chemical Engineering* 32(12): 3198–3212. https://doi.org/10.1016/j.compchemeng.2008.05.010

MultiQuip Inc. 2011. *Soil compaction handbook.* Multiqiup, Inc., Carson, CA.

Paliwal, M.; Kumar, U. 2009. Neural networks and statistical techniques: a review of applications, *Expert Systems with Applications* 36(1): 2–17. https://doi.org/10.1016/j.eswa.2007.10.005

Rogers, M.; Duffy, A. 2012. *Engineering project appraisal.* 2nd ed. John Wiley & Sons.

Shehab, T.; Farooq, M. 2013. Neural network cost estimating model for utility rehabilitation projects, *Engineering, Construction and Architectural Management* 20(2): 118–126. https://doi.org/10.1108/09699981311302991

Shehab, T.; Farooq, M.; Sandhu, S.; Nguyen, T.; Nasr, E. 2010. Cost estimating models for utility rehabilitation projects: neural networks versus regression, *Journal of Pipeline Systems Engineering and Practice* 1(3): 104–110. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000058

Smith, A. E.; Mason, A. K. 2010. Cost estimation predictive modeling: regression versus neural network, *The Engineering Economist: A Journal Devoted to the Problems of Capital Investment* 42(2): 137–161.

Smith, G. N. 1986. *Probability and statistics in civil engineering*. London: Collins.

Sodikov, J. 2005. Cost estimation of highway projects in developing countries:artificial neural network approach, *Journal of the Eastern Asia Society for Transportation Studies* 6: 1036–1047.

Tatari, O.; Kucukvar, M. 2011. Cost premium prediction of certified green buildings: a neural network approach, *Building and Environment* 46(5): 1081–1086. https://doi.org/10.1016/j.buildenv.2010.11.009

Walker, M. 2001. *Introduction to genetic programming*. Tech. Np: University of Montana.

Wideman, R. 1995. *Cost control of capital projects and the project cost management system requirements*. Richmond, BC, Canada: BiTech Publishers Ltd.

Yaghouby, F.; Ayatollahi, A.; Bahramali, R.; Yaghouby, M.; Alavi, A. H. 2010. Towards automatic detection of atrial fibrillation: a hybrid computational approach, *Computers in Biology and Medicine* 40(11–12): 919–930. https://doi.org/10.1016/j.compbiomed.2010.10.004

Yaghouby, F.; Ayatollahi, A.; Bahramali, R.; Yaghouby, M. 2012. Robust genetic programming-based detection of atrial fibrillation using RR intervals, *Expert Systems* 29(2): 183–199.

Neda SHAHRARA. She is a Civil Engineer at the City of San Diego, California, USA. She holds a PhD degree in Construction Management. Her research interests include Public Private Partnership (PPP), investment appraisal, project cost estimation and life cycle costing.

Tahir ÇELIK. He is a professor in the Civil Engineering Department at Cyprus International University, North Cyprus. He is the founder and the director of Construction Engineering and Management Program. His research interests include quality management, life cycle costing, estimating construction projects, construction planning and construction techniques.

Amir H. GANDOMI. He received his PhD in Civil Engineering from University of Akron, OH. He was selected as an elite in 2008 by National Elites Foundation. He used to be a lecturer in several universities and he is currently a distinguished research fellow in an NSF center for the study of evolution in action (BEACON) located at Michigan State University, MI. Dr Gandomi has published over one hundred journal papers and four books. He is one of the most cited researchers in civil engineering field (H-index=41). He also served as associate editor, editor and guest editor in several prestigious journals and delivered keynote talks in international conferences. His research interests are artificial intelligence and their applications in engineering modeling and optimization.