Measurement 118 (2018) 164–171

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Building energy consumption forecast using multi-objective genetic programming

Amirhessam Tahmassebi[a], Amir H. Gandomi[b,*]

[a] Department of Scientific Computing, Florida State University, Tallahassee, FL 32306-4120, USA
[b] School of Business, Stevens Institute of Technology, Hoboken, NJ 07030, USA

ARTICLE INFO

Keywords:
Energy performance
Symbolic regression
Genetic programming

ABSTRACT

A multi-objective genetic programming (MOGP) technique with multiple genes is proposed to formulate the energy performance of residential buildings. Here, it is assumed that loads have linear relation in terms of genes. On this basis, an equation is developed by MOGP method to predict both heating and cooling loads. The proposed evolutionary approach optimizes the most significant predictor input variables in the model for both accuracy and complexity, while simultaneously solving the unknown parameters of the model. In the proposed energy performance model, relative compactness has the most and orientation the least contribution. The proposed MOGP model is simple and has a high degree of accuracy. The results show that MOGP is a suitable tool to generate solid models for complex nonlinear systems with capability of solving big data problems via parallel algorithms.

1. Introduction

Lawrence Livermore National Lab (LLNL), in 2012, has published an annual report on national energy waste that factors energy wasted in everything. They have found that Americans spend $130 billion a year on wasted energy which is the reason of a jump from 70% to 110% in the household income spent on utilities since 2001. The amount of energy wasted by 75,000 average American homes in a single year is equal to the waste that occurred in the 2010 BP Oil Spill MacEachern et al. [20]. It should be noted that roughly 40% of the global energy is consumed by the building sectors Zhu et al. [39]. This issue rises big demand to conserve and control of energy use in buildings.

Baird et al. [2] have employed ample strategies and techniques of energy management, successful conservation programs, energy prediction methods, factors affecting energy consumption, and the essential principles of building energy performance standards to propose a novel framework for classifying the roles and concerns of all groups involved in building energy.

Over the past five decades, a wide variety of categories to improve energy efficiency and sustainability of buildings such as heating, ventilation and air conditioning (HVAC) systems, HVAC equipments, phase change materials (PCM), thermal energy storage (TES), ventilation and multi-zone airflow, zone loads, renewable energy systems, modeling features, and optimized shape of structures have been proposed MacEachern et al. [20], Zhu et al. [39], Baird et al. [2], Tsanas and Xifara [36], Menezes et al. [21], Crawley et al. [5].

For instance, based on structural types of buildings (lightweight or heavyweight), thermal mass can be increased by combining the PCMs and the coating affairs such as gypsum wallboard, concrete, and plaster. Consequently, the differential pricing system for energy can be improved by employing TES which allows us to store the thermal energy (heat or cool) temporarily for later use whenever there is a mismatch between energy supply and demand Zhu et al. [39]. Moreover, modeling features to construct more energy efficient buildings would ensure the construction industry to overcome the aforementioned challenges. However, it has been shown that buildings are not performing as well as expected and there is a performance gap between the predicted energy performance and the actual measured energy Menezes et al. [21], De Wilde [8]. This is known as *performance gap*. The deviation between predicted energy performance based on machine learning algorithms and actual measurements has been mentioned as one of the main types of performance gap De Wilde [8].

Tsanas and Xifara [36] have proposed an accurate quantitative study using statistical machine learning to estimate energy performance of buildings. In this study, they associated the strength of each predictor input variable with statistical metrics using the Spearman rank correlation Zar [38] along with iteratively re-weighted least squares (IRLS) model.

The idea of adjusting weights in the coefficients of the classical regression schemes iteratively can be extended to employ evolutionary

* Corresponding author.
E-mail address: a.h.gandomi@stevens.edu (A.H. Gandomi).

https://doi.org/10.1016/j.measurement.2018.01.032
Received 12 October 2017; Received in revised form 19 January 2018; Accepted 20 January 2018

algorithms to produce a model based on non-Gaussian data with less effect of outliers. Gandomi and Roke [10] have previously designed a model to predict seismic response in structural systems with help of evolutionary algorithms. Genetic programming (GP) Koza [18] is used to predict the statistical parameters of roof drift response under the design basis earthquake using the most effective mechanical and geometric parameters such as rocking behavior, post-tensioning bars, and energy dissipation elements.

In addition to this, Alavi and Gandomi [1] presented promising variants of GP to produce energy-based numerical models for assessment of soil liquefaction. In this regard, linear genetic programming (LGP) Brameier and Banzhaf [4] and multi expression programming (MEP) Oltean and Dumitrescu [23] were employed to conduct relationships between energy density and the factors affecting the capacity energy. Moreover, Muduli and Das [22] were employed multi-gene genetic programming to model uncertainty of SPT-based method for evaluation of seismic soil liquefaction potential. Gandomi et al. [12] have also proposed a novel genetic-based simulated annealing (GSA) non-linear model to formulate hysteretic energy based on various parameters such as earthquake intensity, number of stories, soil type, period, strength index, and energy imparted to the structure.

Many studies in the research area of energy performance have failed to meticulously model the data due to the rigid simplifying mathematical assumptions relying on linear correlations and classical least squares regression techniques Tsanas and Xifara [36]. This suggests to employ the flexibility of GP models in combination with data reduction methods to span new subspaces in which predictor variables are not correlated. Tahmassebi et al. [32–35] have combined GP models with six different data reduction algorithms including, topographic ICA, fast ICA, supervised SVD, SVD, kernel PCA, and PCA in fMRI big data classification problem with roughly 240,000 input predictor variables. This opens new avenues in various scientific fields including energy performance and makes the solutions to big data problems more feasible. Additionally, Vijayaraghavan et al. [37] have employed GP as an optimization approach for quantitative analysis of the data obtained from finite element analysis (FEA) in fracture mechanics modeling of lithium-ion batteries under pinch torsion test Rajan et al. [24].

This study aims to propose a *multi-objective genetic programming (MOGP)* method with multiple genes to formulate the energy performance of residential buildings. The MOGP can automatically select the most substantial predictor input variables in the model, formulate the model structure, and solve the unknown parameters of the regression equation, while simultaneously optimizing for both accuracy and complexity. Here, it is assumed that loads have linear relation in terms of genes. On this basis, an equation is developed by MOGP method to predict both heating and cooling loads. In addition to this, MOGP is written using parallel processing algorithms to accelerate the process Gandomi et al. [13]. This would decrease the run-time of the model which is always substantial in evolutionary algorithms.

2. Genetic Programming (GP)

In 1992, Koza [18] introduced GP as a symbolic optimization technique based on genetic algorithms (GA). The major ability of GP is evolving computer programs based on the Darwins evolution theory. Both GP and GA have been widely used in various optimization problems to find a global optimum solution for a set of predictor input variables. GP does not require any predefined structures of the solution to produce optimum solution to optimize the problem.

In contrast, GA employs a binary encoded version of all predictor input variables to produce a string of numbers as the output of the optimization problem Kwasnicka and Przewozniczek [19]. In other words, GP at first produces a pool of possible solutions also known as population stochastically based on the terminal nodes. Then, the solutions compete with each other at each generation and based on the termination criteria the evolutionary computation stops. Thus, the solution evolves through new generations which are created by genetic operators such as crossover, and mutation. Crossover selects a node from the parental individuals stochastically and replaces the subtree under the selected nodes. In addition to this, mutation generates a node stochastically and truncates and exchanges another node of a tree with the generated node Gandomi et al. [13], Soleimani et al. [30], Garg et al. [15].

To see the performance of the executable program, we need to determine a metric. Likewise machine learning in which we use *loss*, *error*, or *score* as a metric, in GP *fitness* score needs to be optimized as a metric in order to be able to select the best program out of population of individuals Bishop [3]. The higher fitness score leads to a higher probability of winning for an individual in each tournament at each generation. In this regard, the best fitted solution would be found through GP modifications of the individual solution of a population through generations Gandomi et al. [13], Tahmassebi and Gandomi [31]. Despite GP which is a population-based algorithm, most of the machine learning algorithms such as artificial neural networks (ANN), and support vector regression (SVR) are trajectory-based algorithms. Trajectory-based algorithms select a single solution through the learning process while GP deals with a pool of solutions at each generation.

> In words of Darwin [6], owing to this struggle for life, any variation, however slight and from whatever cause proceeding, if it be in any degree profitable to an individual of any species, in its infinitely complex relations to other organic beings and to external nature, will tend to the preservation of that individual, and will generally be inherited by its offspring. The offspring, also, will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive.

3. Problem formulation

Symbolic mathematical regression can be implemented by an evolution of a population of genes using a robust variant of GP, multi-gene genetic programming (MGGP) Gandomi and Alavi [11], Garg et al. [14], Muduli and Das [22]. A typical MGGP model is a weighted linear combination of each gene which contains non-linear terms with respect to the weight coefficients. The general formulation of an MGGP model can be expressed as follows:

$$\hat{y}\,(x,d,\theta) = d_0 + \sum_{i=1}^{n} d_i \cdot G_i(\theta,x) \tag{1}$$

where x is the predictor input matrix, θ is the vector of the unknown parameters for each gene, n is the number of genes, d_0 is a bias term, d_i is the gene weight, and $G_i(\theta,x)$ is the vector of outputs from the i^{th} gene comprising a multi-gene individual. Fig. 1 presents a typical formulation of MGGP model with three input variables $\{x_1,x_2,x_3\}$. However, each gene contains nonlinear terms such as cos and ln, the final model is weighted linear combination of each gene via coefficients $\{d_0,d_1,d_2\}$. Although MGGP is one of the most popular GP variants, it still has limitations for producing models that over-fitting on the testing data. In principle, the underlying relationships of the entire database were not learned, which might give falsify information about the process. Two specific reasons for over-fitting in MGGP models are inappropriate procedure of formation of MGGP model and inappropriate procedure in model selection Garg and Tai [16]. Additional obstacle we might encounter in optimization using MGGP is ending up with an excessively complex developed model since the traditional tree-based GP employs only one objective in the model to maximize the fitness score during the training process. In addition to this, MGGP develops the model in an evolutionary process where the size and shape of the solutions depend on the evolution. This rises a new challenge to add the complexity of the developed model as another objective in optimization problem. The combination of these two objectives is the basis of the multi-objective

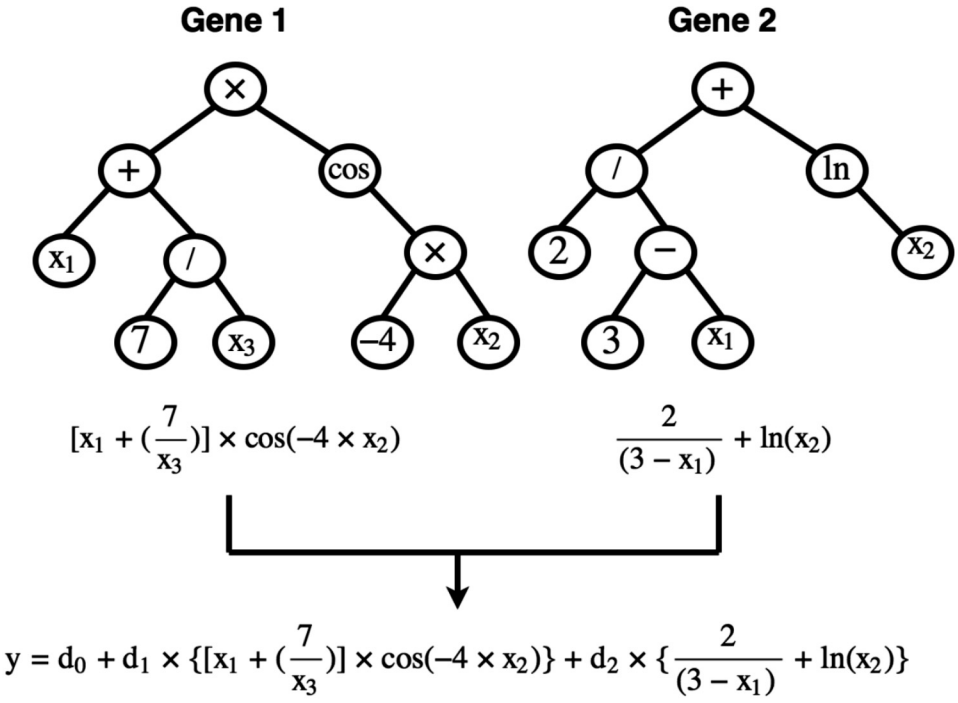

Fig. 1. A typical formulation of MGGP model with three predictor inputs.

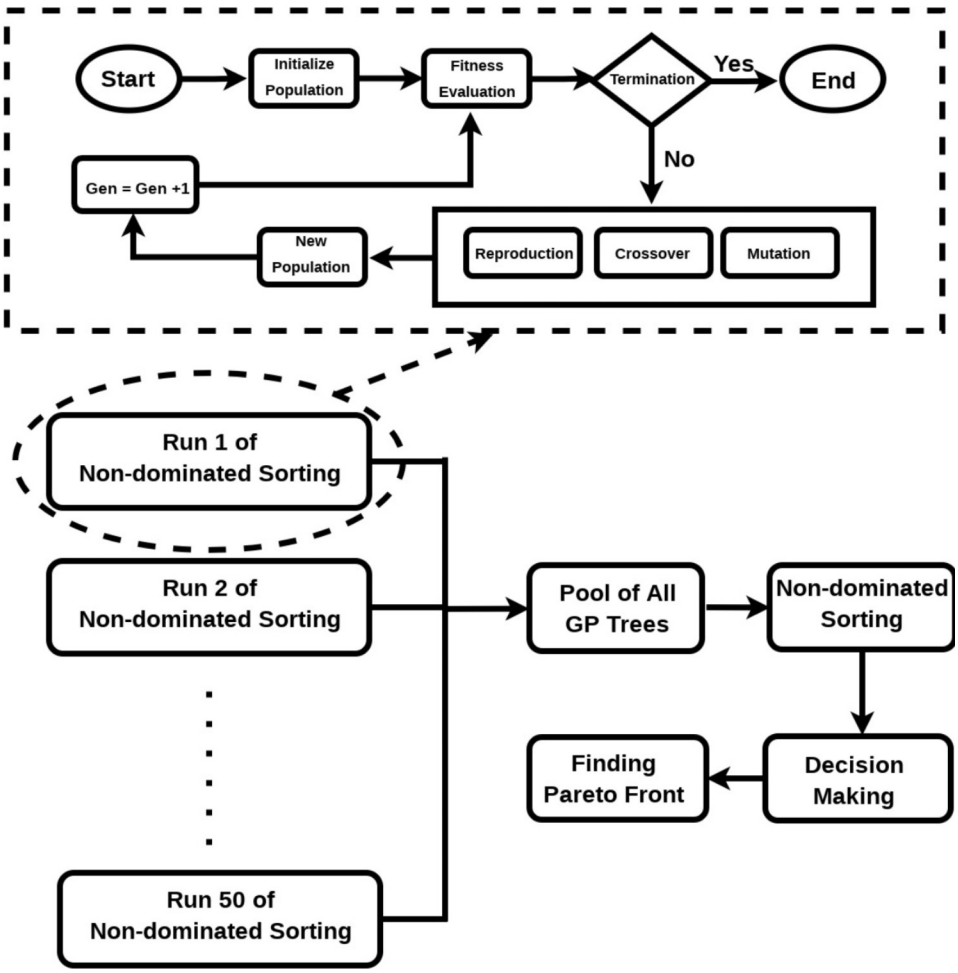

Fig. 2. Flowchart of MOGP.

Table 1
Parameter settings for MOGP algorithm.

Parameter	Settings
Function set	+ ,−,×,/,⌣,log,sin,cos,tanh
Population size	500
Number of generations	1000
Tournament size	10% of population size
Crossover events	0.85
High-level crossover	0.2
Low-level crossover	0.8
Mutation events	0.12
Subtree mutation	0.9
Direct reproduction	0.05

genetic programming (MOGP) algorithm which has roots in multi-objective optimization (MOO) in which finding the optimal solution with respect to multiple fitness functions is desired Slowik and Slowik [29], Słowik and Białko [28]. As it is shown in Fig. 2, each of the 50 runs is based on non-dominated sorting and, therefore, each individual run is MOO. Also, this method creates a pool of non-dominated models from all 50 runs results and does another non-dominated sorting on the pool models to find the final non-dominated models (Pareto front models).

MOGP is an extension of standard GP algorithms which synchronizes the process of maximizing the fitness function and minimizing the complexity of the model. This paper presents the performance of the combination of MGGP and MOGP to improve parsimony and accuracy of the model to predict the energy performance of buildings based on the presented database. To conduct the MOGP algorithm, a pipeline using the GPTIPS 2 toolbox Searson et al. [27], Searson [26] along with scripts written in Python and MATLAB is proposed. In the proposed pipeline, a non-dominated sorting algorithm Deb et al. [9] is employed at the end of each generation of the MOGP model to sort the non-dominant solutions based on their complexity and accuracy. The schematic flowchart of parallel processing in the GP process is shown in Fig. 2. Once the pool of solutions produced by MOGP is ready after 50 runs, the non-dominated sorting algorithm is applied to classify the

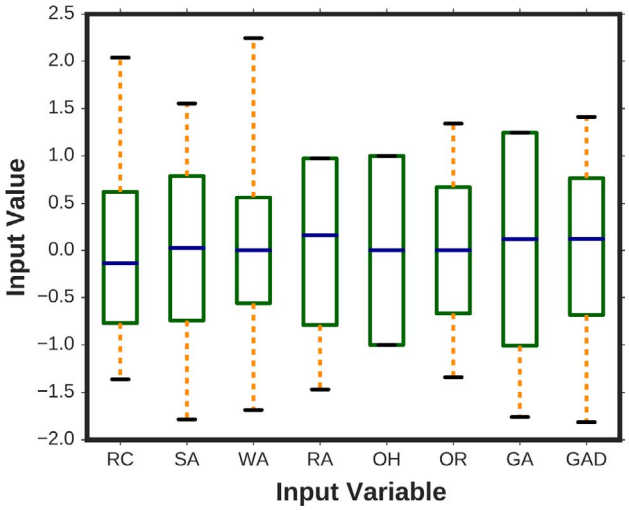

Fig. 4. Box plot presentation of the predictor input variables.

individuals and find Pareto fronts. Then, based on the predefined threshold settings the decision to select the most optimized solution is made.

A set of Pareto optimal solutions that are not dominated by any other solutions present the Pareto front level 1. The solutions that bear Pareto front level 2 are not dominated by any other solutions, apart from those in Pareto front level 1, and so on. Next, a *crowding factor*, such as the average distance of a solution from the nearest solutions on the same Pareto front is calculated for each individual. It would increase the diversity of the population and also giving lower priority to the solutions that are crowded together during the ranking process. Finally, the solutions are ranked according to their position: the solutions on level 1 are ranked above the solutions on level 2, and so on. It should be noted that the solutions that are on the same Pareto front are ranked according to their crowding factors. The top 50% of the population will survive to the next generation, while the rest are eliminated

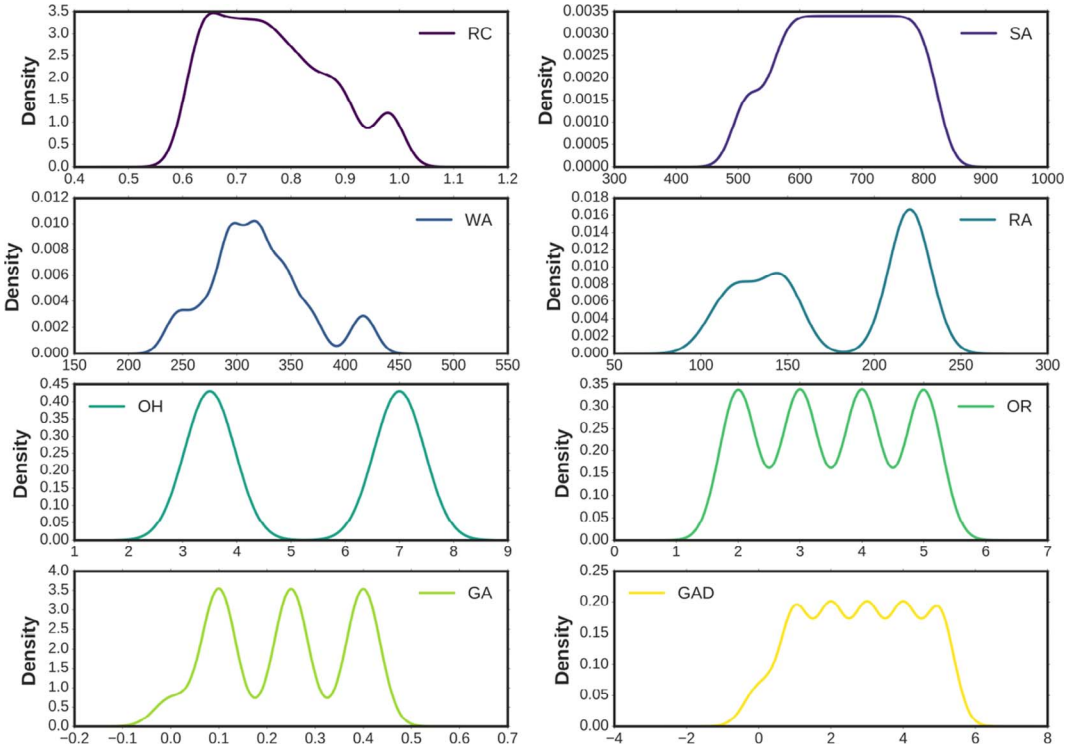

Fig. 3. Density plots of the predictor input variables.

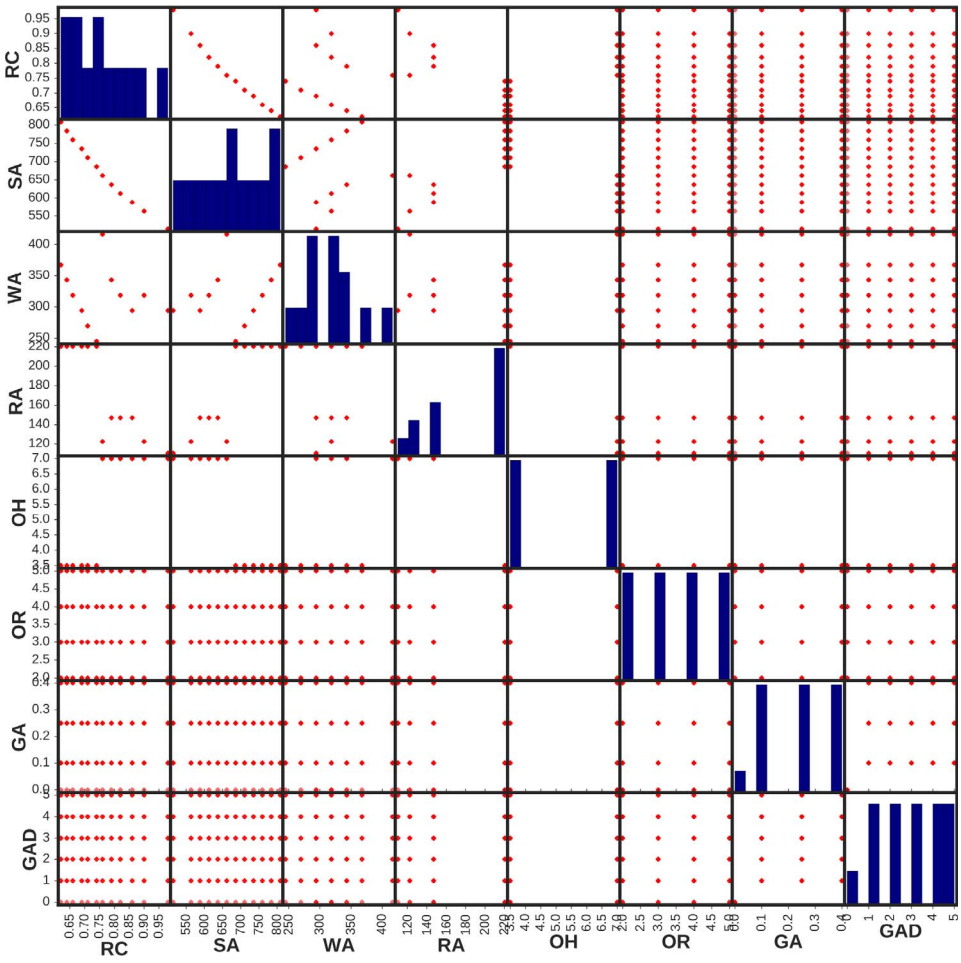

Fig. 5. Scatter matrix presentation of the predictor input variables with their probability histograms.

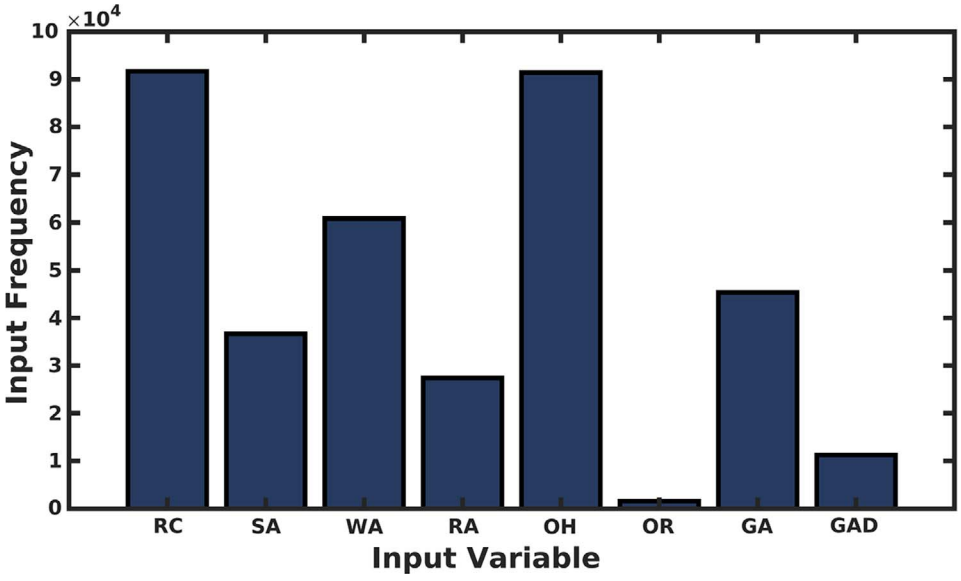

Fig. 6. Histograms of the frequency of each predictor variables used in the best MOGP models as inputs.

Gandomi et al. [13], Searson et al. [27], Searson [26]. Table 1 presents the parameter settings used in the proposed MOGP model.

Next, a database of 768 building samples generated by 12 building forms with volume of 771.75 m^3 using Autodesk Ecotect Analysis have been considered Tsanas and Xifara [36]. Each of the building samples can be characterized by eight predictor input parameters: relative compactness (RC), surface area (SA), wall area (WA), roof area (RA), overall height (OH), orientation (OR), glazing area (GA), and glazing area distribution (GAD). Moreover, two output as heating load (HL) and cooling load (CL) have been recorded for each of the building samples during the simulation process. Fig. 3 illustrates the kernel density estimation (KDE) of each of the eight predictor input variables. KDE is a non-parametric way to estimate the probability density function to smooth finite data sample based on the inferences about the population.

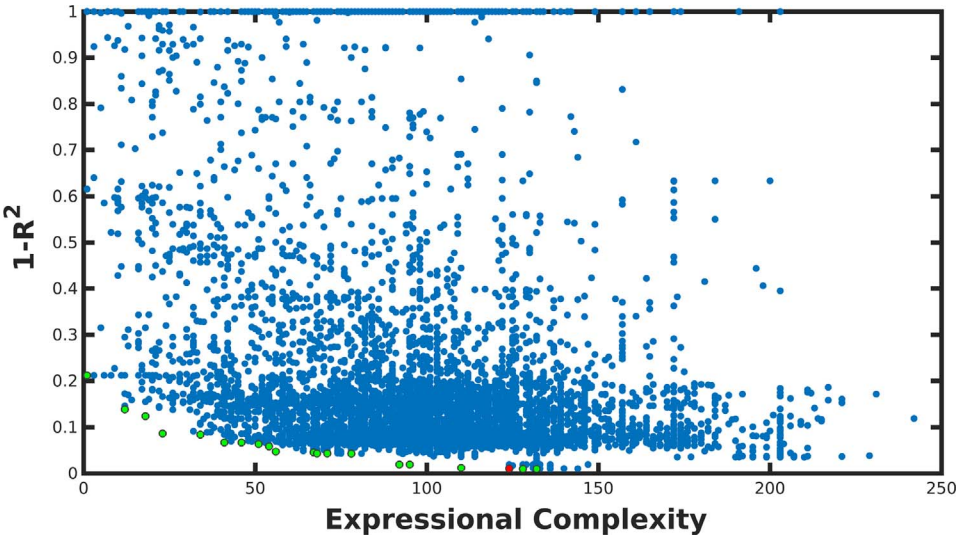

Fig. 7. All models developed using MOGP (solid blue circles), Pareto front results obtained by non-dominated sorting algorithm (solid green circle), and the selected MOGP model (solid red circle). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Regression coefficients for the heating and cooling loads formulas.

Load	Coefficients			
	c_0	c_1	c_2	c_3
Y_{HL}	−22.33	2.70	3.61	−338018.13
Y_{CL}	−3.18	1.97	3.27	−444769.39

Table 3
Statistical parameters of the MOGP models for the external validation.

	Condition	Y_{HL}	Y_{CL}
R	$R \geqslant 0.8$	0.9939	0.9745
$K = \frac{\sum_{i=1}^{n} h_i t_i}{h_i^2}$	$0.85 < K < 1.15$	1.0082	1.0171
$K' = \frac{\sum_{i=1}^{n} h_i t_i}{t_i^2}$	$0.85 < K' < 1.15$	0.9898	0.9767
$R_m = R^2(1-\sqrt{\lvert R^2-R_0^2\rvert})$	$R_m \geqslant 0.5$	0.8806	0.7412
$R_o^2 = 1-\frac{\sum_{i=1}^{n}(t_i-h_i^0)^2}{\sum_{i=1}^{n}(t_i-\overline{t_i})^2}$	$h_i^0 = K \times t_i$	0.9996	0.9977
$R_o'^2 = 1-\frac{\sum_{i=1}^{n}(h_i-t_i^0)^2}{\sum_{i=1}^{n}(h_i-\overline{h_i})^2}$	$t_i^0 = K' \times h_i$	0.9994	0.9955
$m = \frac{R^2-R_0^2}{R^2}$	$\lvert m\rvert < 0.1$	−0.0119	−0.0507
$n = \frac{R^2-R_0'^2}{R^2}$	$\lvert m\rvert < 0.1$	−0.0117	−0.0484

Furthermore, Fig. 4 is depicted to display the variations of the predictor input variables and their outliers without making any assumptions of the underlying statistical distribution. To present the box plot, all the predictor input variables have been normalized to have mean of zero and standard deviation of one.

4. Results and discussion

An MOGP model with 1000 number of generations and population size of 500 for symbolic regression of eight predictor input variables, namely RC, SA, WA, RA, OH, OR, GA, and GAD for one output HL have been developed using GPTIPS 2 toolbox combined with scripts in Python and MATLAB. Fig. 2 depicts the flowchart of the MOGP model. Then, by changing the coefficients of MGGP model, $\{d_i\}$, based on Eq. (1), the regression task has been repeated for the second output, CL. By employing this way, we have proposed an explicit regression model with saving much more runtime without loosing the accuracy. The regression model is as follows:

$$Y = c_0 + c_1\left[\frac{x_5 + 5.92}{\tanh x_1} + x_5\sqrt{x_7}\right] + c_2\left[\frac{x_1 + x_5 + 1.36}{0.58x_3 - x_4 + x_5 - x_5^2}\right] + c_3\left[\frac{x_3}{x_2^2 x_5(x_5 + \tanh x_1 + \sqrt{x_7})}\right] \quad (2)$$

As you can see, the derived Eq. (2) by the MOGP with multiple genes has the same structure for both heating and cooling loads by presenting different weights due to the correlation between heating load and cooling load as the outputs of the regression models. Previously Tsanas and Xifara [36] have explored the statistical relationship among the input values in the database presented in this paper. The presented formulas compared with their formulas which were weighted linear combinations of inputs, have non-linear structure and more complex terms. That would be the reason that they ended up with the mean absolute error of 50% in prediction. In contrast, the proposed MGGP model is a weighted linear combination of each gene which contains non-linear terms.

Fig. 5 shows the scatter matrix to demonstrate the correlation between the eight predictor input variables along with their histograms. It illustrates the multivariate statistics combined with the frequency of the eight input variables. This suggests that classical linear regression models might fail to find an optimized model which comprises all the predictor input variables. Fig. 5 intuitively confirms the importance of using non-linear algorithms to find an accurate and optimized mapping among the input variables and outputs. For instance, it is obvious that RC is inversely proportional to SA. This might be due to the assumptions have been made to generate the database Tsanas and Xifara [36].

As mentioned in Section 3, the final MOGP model formulation portrays the best prediction accuracy along with less complexity. The prediction accuracy has been measured by the coefficient of determination R and the model complexity has been measured by the number of predictor input variables. Fig. 6 demonstrates the frequency of each of the predictor input variables which are used in the best MOGP models. It also pictures that the orientation of the building models has the least contributions in the best MOGP models.

It should be noted that in the MOGP pipeline, we set that the models with $R \geqslant 0.8$ are the best generated models. We executed the code for 50 runs with a population of 500 individuals. In other words, we would come up with 25,000 programs as the pool of solutions as we have shown in Fig. 2 previously. Fig. 7 demonstrates all the models developed using MOGP, Pareto front results obtained by non-dominated sorting algorithm, and the selected MOGP model. Among 25,000 generated MOGP models, 22,159 models had $R \geqslant 0.8$ conditions.

Golbraikh and Tropsha [17] suggested new criteria for external verification of a proposed model: the slope (K or K') of the regression line between the actual data (h_i) and the predicted data (t_i) should be close to 1, and the performance indices $|m|$ and $|n|$ should be lower than 0.1. Recently, Roy and Roy [25] introduced an index (R_m) for external predictability evaluation of models. Their validation criterion is satisfied for $R_m \geqslant 0.5$. The external validation criteria results for the developed models are shown in Table 3. The derived MOGP models satisfy all the proposed conditions.

Based on the presented model by Tsanas and Xifara [36], RC, WA, and RA are the most correlated input variables with respect to the outputs, HL and CL. On the contrary, in the proposed MOGP model RA has not so much contributions in the derived model. To address the difference it should be noted that Tsanas and Xifara [36] have employed IRLS model to overcome the nonlinearity of the database. The IRLS in comparison to GP has less power to find an optimized solution that fits all the inputs with high accuracy and low complexity. In addition to this, the high impacts of outliers as shown in Fig. 4 would be the reason of IRLS's failure to predict the heating and cooling loads with high accuracy Das and Basudhar [7]. It can be also shown by referring to Fig. 3 as the kernel density estimations of the input variables, this conclusion can be drawn that the available linear techniques could not find an appropriate fit for this database in this application.

5. Summary and conclusion

This paper proposes a novel and explicit formulation of building energy consumption forecast via a multi-objective genetic programming (MOGP) technique. The developed MOGP can automatically select the most significant predictor input variables in the model, formulate the model structure, and explicitly solve the unknown parameters of the regression equation through evolution. In addition to this, the derived MOGP optimizes the model for both accuracy and complexity while solving the regression equation. Since the model is coded using parallel algorithms, it can be applied to big data problems as well. To see the performance of the proposed MOGP model, a database of 768 building samples generated by 12 building forms with volume of 771.75 m^3 have been considered. Each of the building samples can be characterized by eight predictor input parameters. After conducting MOGP, the optimum model is selected from Pareto front results with respect to a trade-off between the accuracy and the model complexity. Based on the results, the following conclusions are drawn:

1. The proposed MOGP model shows an explicit non-linear formulation of building energy consumption forecast which correlates the predictor input data with heating load for 99% and cooling load for 97%.
2. By changing the weights presented in Table 2 for the derived MOGP model, the heating and cooling loads (Eq. (2)) would be found.
3. The results show that the relative compactness has the most significant contributions in the best generated MOGP models. On the other hand, the orientation has the least contribution in the derived models respectively.
4. The MOGP algorithm is shown to be a fast enough tool to handle big data problems, and it can be employed to generate solid and accurate models for complex nonlinear systems.
5. The statistical parameters presented in Table 3 prove the validation of the proposed MOGP model which outperforms the previous presented results based on the same database.

Disclosure statement

None.

Acknowledgments

The authors would like to thank Eitan Lees for the careful revision of the final version of the manuscript.

References

[1] A.H. Alavi, A.H. Gandomi, Energy-based numerical models for assessment of soil liquefaction, Geosci. Front. 3 (2012) 541–555.
[2] G. Baird, C. Aun, W. Brauder, M.R. Donn, F. Pool, Energy performance of buildings, 1984.
[3] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[4] M.F. Brameier, W. Banzhaf, Linear Genetic Programming, Springer Science & Business Media, 2007.
[5] D.B. Crawley, J.W. Hand, M. Kummert, B.T. Griffith, Contrasting the capabilities of building energy performance simulation programs, Build. Environ. 43 (2008) 661–673.
[6] C. Darwin, On the Origin of Species by Means of Natural Selection. 1859, Murray Google Scholar, London, 1968.
[7] S.K. Das, P.K. Basudhar, Comparison of intact rock failure criteria using various statistical methods, Acta Geotech. 4 (2009) 223–231.
[8] P. De Wilde, The gap between predicted and measured energy performance of buildings: a framework for investigation, Autom. Constr. 41 (2014) 40–49.
[9] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. 6 (2002) 182–197.
[10] A. Gandomi, D. Roke, Seismic response prediction of self-centering, concentrically-braced frames using genetic programming, in: Structures Congress 2014, 2014, pp. 1221–1232.
[11] A.H. Gandomi, A.H. Alavi, A new multi-gene genetic programming approach to nonlinear system modeling. Part I: materials and structural engineering problems, Neural Comput. Appl. 21 (2012) 171–187.
[12] A.H. Gandomi, A.H. Alavi, A. Asghari, H. Niroomand, A.M. Nazar, An innovative approach for modeling of hysteretic energy demand in steel moment resisting frames, Neural Comput. Appl. 24 (2014) 1285–1291.
[13] A.H. Gandomi, S. Sajedi, B. Kiani, Q. Huang, Genetic programming for experimental big data mining: a case study on concrete creep formulation, Autom. Constr. 70 (2016) 89–97.
[14] A. Garg, A. Garg, K. Tai, S. Sreedeep, An integrated SRM-multi-gene genetic programming approach for prediction of factor of safety of 3-d soil nailed slopes, Eng. Appl. Artif. Intell. 30 (2014) 30–40.
[15] A. Garg, J. Li, J. Hou, C. Berretta, A. Garg, A new computational approach for estimation of wilting point for green infrastructure, Measurement 111 (2017) 351–358.
[16] A. Garg, K. Tai, An improved multi-gene genetic programming approach for the evolution of generalized model in modelling of rapid prototyping process, International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, 2014, pp. 218–226.
[17] A. Golbraikh, A. Tropsha, Beware of q 2!, J. Mol. Graph. Model. 20 (2002) 269–276.
[18] J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection vol. 1, MIT Press, 1992.
[19] H. Kwasnicka, M. Przewozniczek, Multi population pattern searching algorithm: a new evolutionary method based on the idea of messy genetic algorithm, IEEE Trans. Evol. Comput. 15 (2011) 715–734.
[20] A. MacEachern, K. Riling, M. Thompson, Team Earthship, 2015.
[21] A.C. Menezes, A. Cripps, D. Bouchlaghem, R. Buswell, Predicted vs. actual energy performance of non-domestic buildings: using post-occupancy evaluation data to reduce the performance gap, Appl. Energy 97 (2012) 355–364.
[22] P.K. Muduli, S.K. Das, Model uncertainty of SPT-based method for evaluation of seismic soil liquefaction potential using multi-gene genetic programming, Soils Found. 55 (2015) 258–275.
[23] M. Oltean, D. Dumitrescu, Multi Expression Programming, Kluwer, 2002.
[24] A. Rajan, V. Vijayaraghavan, M.P.-L. Ooi, A. Garg, Y.C. Kuang, A simulation-based probabilistic framework for lithium-ion battery modelling, Measurement 115 (2018) 87–94.
[25] P.P. Roy, K. Roy, On some aspects of variable selection for partial least squares regression models, Mol. Inform. 27 (2008) 302–313.
[26] D.P. Searson, GPTIPS 2: an open-source software platform for symbolic data mining, Handbook of Genetic Programming Applications, Springer, 2015, pp. 551–573.
[27] D.P. Searson, D.E. Leahy, M.J. Willis, GPTIPS: an open source genetic programming toolbox for multigene symbolic regression, in: Proceedings of the International Multiconference of Engineers and Computer Scientists, Citeseer, vol. 1, 2010, pp. 77–80.
[28] A. Słowik, M. Białko, Design and multi-objective optimization of combinational digital circuits using evolutionary algorithm with multi-layer chromosomes, in: Artificial Intelligence and Soft Computing–ICAISC 2008, 2008, pp. 479–488.
[29] A. Slowik, J. Slowik, Multi-objective optimization of surface grinding process with the use of evolutionary algorithm with remembered pareto set, Int. J. Adv. Manuf. Technol. 37 (2008) 657–669.
[30] S. Soleimani, S. Rajaei, P. Jiao, A. Sabz, S. Soheilinia, New prediction models for unconfined compressive strength of geopolymer stabilized soil using multi-gen genetic programming, Measurement 113 (2018) 99–107.
[31] A. Tahmassebi, A.H. Gandomi, Genetic programming based on error decomposition: a big data approach, Genetic Programming Theory and Practice XIV, Springer, 2017.

[32] A. Tahmassebi, A.H. Gandomi, I. McCann, M. Schulte, L. Schmaal, A.E. Goudriaan, A. Meyer-Baese, fMRI smoking cessation classification using genetic programming, in: Workshop on Data Science meets Optimization, 2017a.
[33] A. Tahmassebi, A.H. Gandomi, I. McCann, M.H. Schulte, L. Schmaal, A.E. Goudriaan, A. Meyer-Baese, An evolutionary approach for fMRI big data classification, in: IEEE Congress on Evolutionary Computation, 2017b.
[34] A. Tahmassebi, A.H. Gandomi, A. Meyer-Bäse, High performance gp-based approach for fMRI big data classification, Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact PEARC17, ACM, New York, NY, USA, 2017, pp. 57:1–57:4, , http://dx.doi.org/10.1145/3093338.3104145 <http://doi.acm.org/10.1145/3093338.3104145>.
[35] A. Tahmassebi, A.H. Gandomi, M.H. Schulte, I. McCann, L. Schmaal, A.E. Goudriaan, A. Meyer-Baese, fMRI smoking cessation classification, IEEE Trans. Cybernet. (2017d).
[36] A. Tsanas, A. Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy Build. 49 (2012) 560–567.
[37] V. Vijayaraghavan, A. Garg, L. Gao, Fracture mechanics modelling of lithium-ion batteries under pinch torsion test, Measurement 114 (2018) 382–389.
[38] J.H. Zar, Spearman rank correlation, Encycl. Biostat. (1998).
[39] N. Zhu, Z. Ma, S. Wang, Dynamic characteristics and energy performance of buildings using phase change materials: a review, Energy Convers. Manage. 50 (2009) 3169–3181.