Analysis of approximations of the gas compressibility factor derived from genetic algorithms

Olga Kochueva[1,*], and *Vladislav* Zadorozhnyy[2]

[1]National University of Oil and Gas "Gubkin University", Moscow, Russia
[2]In-Pipe Robot Inc., 1000 N West Street STE 1200, Wilmington, Delaware, USA

Abstract. Hydraulic calculations are the primary tool for rational technical decisions related to the design and operation of pipeline systems. The compressibility factor is introduced into the gas equation of state to account for its real properties and depends on the pressure, temperature, and gas component composition. At present, the search for an accurate and computationally efficient approximation for the compressibility factor remains an urgent problem. This paper presents a methodology for constructing an approximation based on symbolic regression, and the proposed dependencies analysis provided. The average relative error of the presented models is 0.03%.

1 Introduction and motivation

For hydraulic calculations of gas flow parameters in the gas pipeline it is necessary to calculate the compressibility factor $z(p,T)$, which allows to take into account the deviation of real gas behavior from the ideal one. Most explicit and implicit approximation dependences have been developed based on the accumulated significant volume of experimental data. The calculation of the compressibility factor with regard to gas fraction composition can be performed according to the equation of state (in recent years, the work of most researchers is based on GERG-2008, AGA8[1,2], AGA10), where the problem is reduced to an iterative method of solving a nonlinear equation, refer to [3] as well. The time of calculations plays a significant role in non-stationary modeling of gas transfer modes in main pipelines as well as determining optimal mode of gas transmission and parameters identification of a gas transportation system. The work aimed to build approximation dependences for the compressibility factor in an explicit form using the genetic programming method (another name is symbolic regression) and to analyze the quality of the resulting models. It is important to note the approaches to solving the problem of approximating the compressibility factor, presented in a number of publications. In [4] the relationship in explicit form, constructed on the basis of a set of experimental data (3038 records) is presented. The correlation is built as a fraction, the numerator and denominator of which are functions containing the reduced pressures and temperature in various degrees

[*] Corresponding author: olgakoch@mail.ru

and their logarithms, there are 20 coefficients in the formula. The paper [5] presents the formula that is a modification of the implicit relation [6]. The resulting model contains 19 coefficients, and the polynomial functions of reduced temperatures and pressure and exponential functions of reciprocal temperature are used in the calculation. A feature of [7] is the division of the range of reduced pressure values into 2 subsets, a separate model is built for each range. Due to this it became feasible to reach a higher accuracy of the model, the group method of data handling (GMDH) was used in the work. The authors note that the models built using GMDH do not include logarithmic or exponential functions and can be easily used in programming. The paper [8] introduces a model based on a multidimensional nonlinear relationship, built on a sample of 6988 values obtained from the digitized Standing–Katz diagram [9,10]. The authors give a formula that is the quotient of two polynomials, the formula has 20 coefficients, an average error (MAPE) about 1.5% flowmeter is indicated. To calculate the compressibility factor in [11], an artificial neural network with two hidden layers is proposed it was trained on 4158 experimental data entries. However, it is difficult to compare the results of mentioned models since the testing was carried out on a disparate data. In many works, authors usually deal with charts determining the ratio between the obtained and experimental values of the compressibility factor, they also provide the determination coefficient values and the average absolute error, but the majority of them do not conduct a complete analysis of the dependence of the error on the input variables values.

2 Methodology

Symbolic regression or genetic programming [12] is one of the methods of machine learning. The advantage of this method is the possibility to obtain the relationship between input and output variables in an analytical form, and there is no need to specify the shape of the function pattern in advance, it will be determined directly during the execution of the genetic algorithm. In contrast to neural networks and other machine learning algorithms, the obtained models are not a "black box" and allow you to analyze the interaction between the input variables. There is a fairly large number of works describing the application of this method to solving problems encountered in practice in a wide variety of subject areas, for instance [13, 14, 15, 16]. The terminology of the method comes from biology. The resulting analytical function is represented as a chromosome (or an individual) and is defined by a set of genes, which may include:

- variables,
- arithmetic operations, including unary minus,
- functions (log, exp, sqrt, ^2, ^3, abs, sin, cos, tanh, etc.),
- constants.

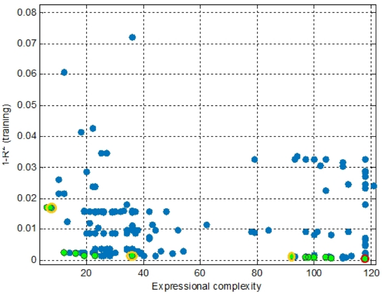
When an initial population is created, a set of chromosomes (functions) will be randomly generated from a set of genes. Speaking about the fitness of the j-th individual (the function which approximates the compressibility factor values) $Z_j(x_1,x_2,...,x_n)$, we consider the sum of squares of deviations of the values calculated by the formula obtained for each set $(x_{1i},x_{2i},...,x_{ni})$ of input variables in the training set from the known values of the compressibility factor z_i. A "-" sign is required, since due to the evolutionary principle, the most adapted individuals survive, and the values of the adaptability function should tend to the maximum.

$$ff_j = -\sum_{i=1}^{n}(Z_j(x_{1i},x_{2i},...,x_{ni}) - z_i)^2 \to \max \tag{1}$$

Further, the selection takes place according to the chosen method. To obtain offspring, it's required to select points where a chromosome splitting can correctly occur, then each offspring receives the left part of the chromosome from one parent and the right one from the other. To avoid a local minimum, a mutation operation is applied. It is performed for descendants with a probability of 0.01-0.05 by randomly changing one of the chromosome genes. The specific probability value is a tuning parameter, as well as the number of population individuals and the selection method. Then a new population is created, where individuals from the parents' generation with the highest fitness function values and descendants are transferred. The condition of procedure termination is the formation of a given number of generations, reaching the assigned limit of the fitness function value, or exceeding the predetermined maximum calculation time. Then the individual with the best fitness function value is selected from the last population. In the problem under consideration, several individuals with high values of the fitness function and different estimates of the computational complexity of the obtained models may be of interest. The authors didn't pursue the goal to obtain universally applicable models hence two pressure ranges: P1 3.5-5.6 MPa и P2 5.5-7.5 MPa, the temperature between 273 and 333 K, the gas compositions as a mixture of methane (from 90% to 97%), hydrocarbons from C_2 to C_6 plus nitrogen, carbon dioxide, and helium were taken to generate initial data. Since many components are contained in natural gas in small amounts, the study planned to analyze the possibility of reducing the number of input variables without compromising the model quality, therefore the molar mass of the gas was added to the input parameters. For different gas compositions, pressure values were set with a step of 0.02MPa, temperature with a step of 2K, and the values of the compressibility factor were calculated with the method [3] (its analog can be found in [2]). Thus, initial data for model training and testing were generated with the number of entries 15900 for the range P1 and 13300 for the range P2. All input variables were normalized in the range [0, 1]. All further models are given for normalized input parameters. The GPTIPS open-source software [17] was used to generate the models.

3 Results and discussion

Consider the results of one of the runs of the genetic algorithm (for the pressure range P1), shown in Fig. 1. Each model is represented as a point with the following coordinates: the value (1-R^2) for the training set is on the vertical axis, the computational complexity of the model is on the horizontal axis. The blue dots show the models of the last population that are not Pareto-optimal. Pareto-front models with the highest R^2 value (with the lowest 1- R^2 value respectively) for a fixed computational complexity are highlighted in green, and exactly these models will be of most interest to the user. Models with high computational complexity may be overtrained, so the analysis should include the Pareto models on the lower left part of the graph (high R^2 and low computational complexity). The table on the right shows the models marked with red dots in Fig. 1.



R^2	Model complexity	Model
0.999	118	$0.0262 \cdot T - 0.0168 \cdot P - 0.00804 \cdot Mm - 0.00122 \cdot T^2 \cdot (Mm+P) - 0.00115 \cdot (2 \cdot Mm + 2 \cdot T) \cdot (Mm+P+2 \cdot T) + 0.00872 \cdot T \cdot (Mm+P) + 0.00114 \cdot Mm \cdot (CO2 + 2 \cdot Methane + N + T) + 0.923$
0.999	92	$0.0262 \cdot T - 0.0169 \cdot P - 0.00844 \cdot Mm - 0.00134 \cdot T^2 \cdot (Mm+P) - 0.00105 \cdot (2 \cdot Mm + 2 \cdot T) \cdot (Mm+P+2 \cdot T) + 0.00843 \cdot T \cdot (Mm+P) + 0.923$
0.998	36	$0.0263 \cdot T - 0.0166 \cdot P - 0.01 \cdot Mm + 0.00639 \cdot P \cdot T + 0.00214 \cdot T \cdot (Mm - P \cdot T) - 0.00488 \cdot T^2 + 0.924$
0.983	6	$0.0265 \cdot T - 0.0176 \cdot P - 0.00998 \cdot Mm - 0.00478 \cdot T^2 + 0.924$

Fig. 1. Pareto front of models in terms of model performance (1 – R^2) and model complexity.

The first column of the table (Fig. 1, right) shows the values of the coefficient of determination R^2, the second column presents an estimate of the computational complexity of the model based on the number of nodes and tree depth, and the third column provides the obtained analytical relationships, where *T*, *P*, *Mm*, *Methane*, *N*, *CO2* are temperature, pressure, molar mass, methane, nitrogen, carbon dioxide, respectively (all the variables are normalized, dimensionless). The simpler formulas show more clearly the influence of the input variables, but their coefficient of determination is lower. We can see that the coefficients with the main influencing variables are quite stable.

The REC (Regression Error Characteristics) [18] are plotted for the test sample (Fig. 2 left), the abscissa axis is the absolute value of deviation of the predicted value *Z* from the actual one, the ordinate axis is the proportion of data for which the prediction error does not exceed *x*. Therefore, REC can be interpreted as an empirical distribution function of the error for the constructed model. From the REC plots in Fig. 2, we can conclude that only the last and simplest model from the table (C6) is significantly inferior in accuracy to the others. According to Fig. 2, at least 98% of models C92 and C36 predictions and all model C118 predictions have an error not exceeding 0.002. A common way to show the quality of the resulting model is to plot the "predicted vs actual" graph (Fig. 2, right).

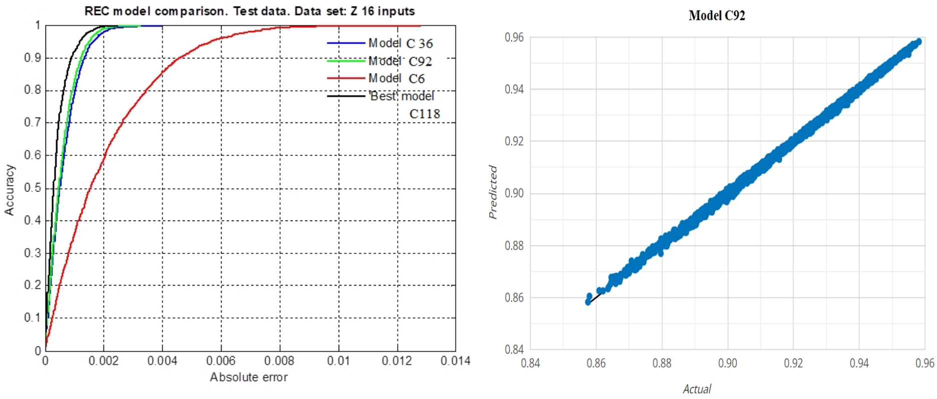


Fig. 2. Model validation (a - REC curves, b - predicted *z* vs actual *z*).

It is worth noting that the error of calculation methods recommended in [1] and [3] for the selected pressure and temperature ranges is estimated within 0.2%, so the three formulas under consideration (models C36, C92 C118) are suitable as approximations for calculating the compressibility factor. One of the indicators of model quality is the absence of a correlation between predicted values and values of input variables. So, it is necessary to plot graphs of calculated values deviations from actual ones (often called "residuals") for the test sample. Figure 3 shows plots of residuals depending on pressure and temperature values for model C36 for the test sample. We can see that the scatter is random.

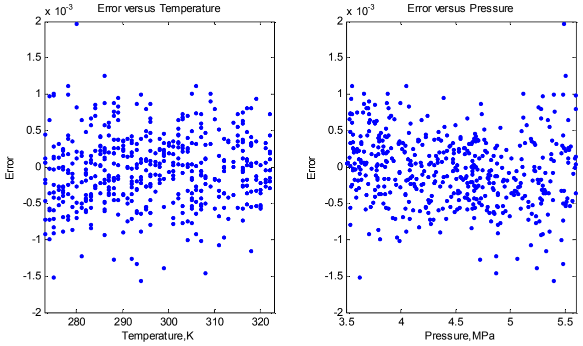


Fig. 3. Model validation (a - REC curves, b- predicted *z* vs actual *z*).

4 Conclusion

The presented approximations for calculating the compressibility factor have R^2=0.999, MAE=0.001, MAPE=0.1%. The error of the calculation methods, by which the input data for model building were obtained, is within 0.2%. The results indicate that the goal of the work has been achieved, and approximations have been obtained that can successfully replace computationally demanding iterative procedures without loss of accuracy. The accuracy of the presented relationships is comparable with models built with Random Forest algorithm and neural networks. The resulting models are easy to integrate into existing software and in higher-level models. The symbolic regression method can be applied both as alternative to computationally complex procedures and when working with real data. The introduced procedure is an example of the so-called surrogate modeling. This is a computer modeling technique where machine learning methods are used to build a fast (surrogate) model that allows you to get a result with acceptable accuracy on data from a complex (physically proven, built on a solution of systems of nonlinear algebraic equations or partial differential equations) and resource-intensive model of an object or process. Subsequently, this model trained on a large amount of calculated data can be applied to replace the original model when analyzing the development of risk situations and searching for optimal gas transportation modes, in software designed for staff training on computer simulators.

References

1. ISO 12213-2:2006,. *Natural Gas — Calculation of Compression Factor* Switzerland, Geneva, ISO, (2006)
2. Repository for the supplementary files to AGA 8, https://pages.nist.gov/AGA8/ NIST USA (accessed on 10.02.2022)
3. GOST R 30319.3-2015. *Natural Gas. Methods for Calculating the Physical Properties* (In Russian), Standardinform, Moscow (2016)
4. N. Azizi, R. Behbahani, M.A. Isazadeh, J. Nat. Gas Chem. **19**, 642–645 (2010)
5. L.A. Kareem, T.M. Iwalewa, M. Al-Marhoun, J. Petrol Explor Prod Technol. **6**, 481–492 (2016)
6. K.R. Hall, L.Yarborough, Oil Gas J. **71**, 82–92 (1973)
7. L. Lin, S. Li, S. Sun, Y. Yuan, M. Yang, Flow Measurement and Instrumentation **71**, 101677 (2020)
8. Y. Wang, J. Ye, Sh. Wu, Energy Reports, 8(2), Supplement 2, 130-137, (2022)
9. M. Standing, D. Katz, Density of natural gases. Trans AIME 1942;146(1):140–9.
10. F. Poettman, P.Carpenter, Drilling and production practice. New York: American Petroleum Institute; 1952, 257–317
11. N. Azizi, M. Rezakazemi, M. Zarei, Neural Computing and Applications. **31(1)**, 55-64 (2019)
12. J.R. Koza *Genetic programming: on the programming of computers by means of natural selection* The MIT Press, Cambridge, 1992
13. A.Gandomi, A. Alavi, Neural Comput & Applic. **21(1)** , 171–187 (2011)
14. P. Praks, D. Brkic, Water. **10**, 1175, (2018)
15. O. Kochueva, K. Nikolskii, Computation **9**, 139 (2021)
16. O. Kochueva, Business magazine Neftegaz.RU, **5–6**, 14-20 (2022)
17. D.P. Searson, D.E. Leahy, M.J. Willis, LNEE, **70**, 83-93 (2011)
18. J. Bi, K. Bennett, *Regression error characteristic curves* (ICML-2003, Washington DC, 2003)