Comparative Analysis of Statistical Models for the Task of Natural Gas Composition Analysis

I. A. Brokarev, S. V. Vaskovskii

Abstract. A large number of statistical methods are being developed to solve the problem of natural gas composition analysis. Statistical models are used in these methods for determination of natural gas composition by its known physical parameters. The choice of a statistical model for the method under discussion is a difficult task. No general algorithm has been found for selecting a model for a specific task. Basic statistical models, that are often used in practice, are studied in the article. The comparative analysis of the models is carried out according to a number of important criteria for solving the discussed problem. As a result, it is concluded that the neural network model is the most effective model for the natural gas composition analysis. Recommendations are given on choosing a statistical model in the tasks of natural gas quality analysis that are similar to the problem under consideration.

Keywords: machine learning, statistical models, neural network analysis, composition analysis, natural gas.

DOI 10.14357/20718632200104

Introduction

Currently, the natural gas composition analysis is an urgent task for the gas industry. A change in the natural gas composition due to technological peculiarities of the transportation and storage processes leads to a change in a number of energy characteristics of the gas (in particular, its calorific value [1]). In its turn, it leads to difficulties in calculating the natural gas cost indicators. The correlation methods are developed for real time natural gas composition analysis for this reason [2; 3]. Various statistical models are often used in correlation methods due to the impossibility or high complexity of solving the problems with traditional methods. Having a number of advantages over traditional algorithms [4], such statistical models assign unknown outputs to the known input values using existing dependencies between them.

The choice of a model for the problem under discussion is mostly made by heuristic methods due to the lack of a general algorithm for choosing a model and a variety of both statistical models and architectures of individual models. The article provides a comparative analysis of the main models that are used to solve the task of natural gas composition analysis. On the basis of this analysis, the conclusions are drawn about a specific model, that is most appropriate to apply to existing data.

1. The methodology of statistical model comparative analysis

A number of preliminary procedures were developed and implemented prior to conducting a comparative analysis of statistical models for natural gas composition determination: selection of initial data and ensuring uniformity of conditions for model training, as well as selecting necessary data for model testing; selection of statistical models for comparison; selection of criteria and characteristics to be used for model comparison.

The first step is to select data for model training and testing. The data must comply with the requirements for the natural gas [5], but should not create unsolvable problems in the process of model analysis using this data.

For the initial data selection a sample of gas mixtures based on the Russian natural gas was simulated taking into account the permissible ranges of the molar fractions of the components by sorting out all possible combinations of components. Then the simulated gas mixtures were reduced to equivalent four-component pseudo-gas mixtures to simplify the analysis of the considered statistical models [6]. Based on the obtained natural gas model, the output parameters for the model are concentrations of effective components of the pseudogas mixture that can be determined, namely, methane, propane, nitrogen and carbon dioxide. The input parameters for the model should be physical parameters of the natural gas. The criteria for choosing input physical parameters are: availability of the applied technology for measuring the parameter; availability of a commercially available and relatively inexpensive parameter measuring technique; a correlation between the gas parameter and the composition; a high correlation between the parameter and the composition and a low correlation with the other input parameters.

Having reviewed known methods for natural gas quality control [2] and taken into account instruments for measuring the necessary physical gas parameters [7-11] the following parameters were selected as possible input parameters for the model: the speed of sound (c), thermal conductivity coefficient (χ), dynamic viscosity (η), dielectric permittivity (ε) and carbon dioxide concentration (X_{CO2}). Methane concentration (X_{CH4}), nitrogen concentration (X_{N2}) and propane concentration (X_{C3H8}) in equivalent four-component pseudo-gas mixture were chosen as the output parameters.

The correlation analysis was performed for selection of input parameters and elimination of their possible multicollinearity. Pearson correlation coefficients are calculated for each pair of the studied parameters. These coefficients can be used to determine a linear relationship between two parameters. In the general case, the Pearson correlation coefficient r for samples of parameters X_1 and X_2 is calculated as follows:

$$r = \frac{\sum_{i=1}^{N}(X_{1i} - \overline{X_1})(X_{2i} - \overline{X_2})}{\sqrt{\sum_{i=1}^{N}(X_{1i} - \overline{X_1})^2 \sum_{i=1}^{N}(X_{2i} - \overline{X_2})^2}}\,.$$

Carbon dioxide concentration, the speed of sound, and thermal conductivity coefficient were selected as input parameters due to the correlation analysis results, that are shown in Table 1.

It should be noted that prior to training the model, the initial data are cross-validated and normalized in order to be able to be used uniformly and improve the forecast results of the studied models. A sample of 552000 elements was used for the task under discussion. The ranges of concentrations included in the training set and their parameters that were used as input parameters are shown in Table 2.

The next step is to get a number of statistical models that can be used to solve the task of the natural gas composition analysis. This choice is

Table 1. Correlation coefficients of input parameters against the gas component concentrations

	c, m/s	χ, W/(m·K)	η, Pa·s	ε, -	X_{CH4}, %	X_{N2}, %	X_{CO2}, %	X_{C3H8}, %
c, m/s	1	0,992	-0,278	-0,720	0,929	-0,243	-0,623	-0,743
χ, W/(m·K)	0,992	1	-0,162	-0,799	0,885	-0,159	-0,564	-0,810
η, Pa·s	-0,278	-0,162	1	-0,461	-0,485	0,516	0,746	-0,421
ε, -	-0,720	-0,799	-0,461	1	-0,481	-0,222	0,083	0,971
X_{CH4}, %	0,929	0,885	-0,485	-0,481	1	-0,577	-0,577	-0,577
X_{N2}, %	-0,243	-0,159	0,516	-0,222	-0,577	1	0	0
X_{CO2}, %	-0,623	-0,564	0,746	0,083	-0,577	0	1	0
X_{C3H8}, %	-0,743	-0,810	-0,421	0,971	-0,577	0	0	1

Table 2. Ranges of gas mixture concentraions and their physical parameters for the training sample

Parameter	Range
X_{CH4}, %	85 – 100
X_{N2}, %	0 – 5
X_{C3H8}, %	0 – 5
X_{CO2}, %	0 – 5
c, m/s	401.49 – 445.02
χ, mW/(m·K)	30.59 – 33.27

based on an analysis of sources that address the problems arising when selecting statistical models for specific tasks of the oil and gas industry [12-15], as well as the practical feasibility of implementing the selected statistical models. The following models were selected for comparative analysis on the basis of the study results:

- multiparameter linear regression;
- ridge regression;
- Gaussian process regression;
- neural network model (multilayer perceptron);
- recurrent neural network model;
- recurrent neural network model with long short-term memory.

The accuracy characteristics of the model are often used as the main parameters to make a conclusion about the possibility of using the statistical model [16]. Various accuracy parameters are calculated for both the training and test samples in the conducted comparative analysis. The fact that the statistical model can show good results on a training set, but a high error on a testing set is taken into account. The time that was spent for the model training is another important parameter in assessing performance of the statistical model. For large samples training of models with complex architecture can take a long time that may not respond to the required characteristics.

The following parameters are calculated to assess model accuracy: mean absolute error (MAE) and mean absolute percentage error (MAPE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | Y_{i\,output} - Y_{i\,target} |$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n} | \frac{Y_{i\,output} - Y_{i\,target}}{Y_{i\,target}} |,$$

where Y_{output} are values that were obtained from the statistical model, Y_{target} are initial target values, n is a test or train sample size.

Taking into account the fact that the model can have a satisfactory average error, but still have outliers at certain points it is also necessary to calculate the maximum absolute error (MaxAE) and maximum absolute percentage error (MaxAPE).

$$MaxAE = \max | Y_{output} - Y_{target} |$$

$$MaxAPE = 100\% \max | \frac{Y_{output} - Y_{target}}{Y_{target}} |,$$

where max is an operation of calculating the maximum value.

2. Statistical models under study

The multiparameter linear regression can be considered in the studied problem as a reference model [17]. It can be used to obtain a result that will be taken to compare accuracy of other models with regression. In case of multiparameter linear regression, the value of Y depends on several independent quantities x_i (i=1…m). The initial points are in an m+1-dimensional space and are approximated by the m-dimensional hyperplane. The system of equations for multiparameter linear regression can be written as:

$$\begin{cases} y_1 = \beta_0 + x_{11}\beta_1 + ... + x_{1m}\beta_m \\ y_2 = \beta_0 + x_{21}\beta_1 + ... + x_{2m}\beta_m \\ ... \\ y_n = \beta_0 + x_{n1}\beta_1 + ... + x_{nm}\beta_m \end{cases}$$

This system of equations can be written in matrix form taking into account the error vector e:

$$Y = X\beta + e$$

where β is a (m+1)-dimensional parameter vector, X is a matrix of row-vectors x_i that can be written as:

$$X = \begin{pmatrix} x_1^T \\ x_1^T \\ . \\ . \\ . \\ x_m^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & ... & x_{1m} \\ 1 & x_{21} & ... & x_{2m} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & x_{n1} & ... & x_{nm} \end{pmatrix}$$

In order to find an estimate of the linear regression parameters, it is necessary to find the minimum of the sum of squared errors, that is written in the matrix form as follows:

$$\sum_i e_i^2 = e^T e = (Y - X\beta)^T (Y - X\beta).$$

To find an estimate of the parameters it is necessary to use the condition that the partial derivative is zero at the minimum point. A system of normal equations for multiparameter linear regression in the matrix form can be obtained by differentiating the expression for the sum of the squared errors with respect to the variable β and equating the resulting partial derivative to zero:

$$(X^T X)\beta = X^T Y.$$

The estimation of the parameters of multiparameter regression is the solution to this system using the least squares method, which may be shown as follows:

$$\overset{\wedge}{\beta} = (X^T X)^{-1} X^T Y.$$

The multiparameter linear regression is trained on the initial data with calculating an estimate of the parameters for each component of the gas mixture under study. Then the model is tested on data that was not involved in the training process. The accuracy estimates for the model are calculated on the basis of the test results.

The ridge regression is used in tasks with data redundancy as one of the methods of dimensionality reduction [18]. In the problem under study, this is possible when the input parameters correlate with each other, i.e. multicollinearity is not completely eliminated by the correlation analysis. Multicollinearity can lead to instability of estimates of regression coefficients and poor conditioning of the X^TX matrix, that leads to instability of the normal linear regression equation solution. The ridge regression method consists in introducing an additional regularizing parameter τ into the minimized functional. The applied regularization makes it possible to reduce the condition number of the X^TX matrix and obtain a more stable solution. The parameters of the regression model with regularization are found through minimizing the functional β^*:

$$\beta^* = \arg\min(\|Y - X\beta\|^2 + \tau\|\beta\|^2).$$

The solution to the minimization problem is found in the same way as to the linear regression:

$$\beta^* = (X^T X + \tau I)^{-1} X^T Y\,,$$

where I is an identity matrix.

Let us consider the condition number μ of the matrix $M = X^TX+\tau E$ which is the ratio of its maximum to the minimum singular number. For the regularized matrix under consideration, it is equal to

$$\mu(M) = \frac{\lambda_1 + \tau}{\lambda_n + \tau},$$

where λ_i are matrix eigenvalues X^TX.

From the above formula it is seen that an increase in the regularization parameter leads to a decrease in the condition number of the regularization matrix. The smaller this parameter, the less is the error of the solution regarding errors in the input data. Moreover, an increase in the regularization parameter leads to a decrease in the norm of the parameter vector. It is worth noting that the ridge regression method improves the stability of the parameters of the regression model, but does not nullify any of them [19].

Ridge regression is trained and tested on the same data division scheme as multiparameter linear regression. Based on the essence of the ridge regression method higher accuracy estimates for this model can be assumed.

The Gaussian process regression [20] is a nonparametric probabilistic model of the process, all finite-dimensional distributions of which are normal. The Gaussian process regression model addresses the question of predicting the value of a response variable, given the new input vector and the training data $\{(x_i ,y_i);\ i=1...n\}$. The Gaussian process regression model explains the response by introducing latent variables, $f(x_i)\,(i=1...n)$ from a Gaussian process, and explicit basis functions. The covariance function of the latent variables captures the smoothness of the response and basis functions project the inputs x into a p-dimensional feature space.

The Gaussian process is defined by the mathematical expectation function and the covariance function. The Gaussian process (GP) is designated as follows:

$$f(x) \sim GP(m(x), k(x, x'))\,,$$

where f(x) is a Gaussian process latent function values, x is a set of the training inputs, ~ is sign that means “distributed according to”, k(x, x′) is a covariance function evaluated at x and x′, m(x) is the mean function of a Gaussian process.

If f(x) is a Gaussian process, then the mathematical expectation function and the covariance function can be represented as follows:

$$m(x) = E(f(x)),$$

$$k(x,x') = E((f(x)-m(x))(f(x')-m(x'))),$$

where E is the symbol of mathematical expectation.

The Gaussian process is a set of random variables, such that any finite number of them have a joint Gaussian distribution. If f(x) is a Gaussian process, then given n observations $x_1 \ldots x_n$, the joint distribution of the random variables $f(x_1)\ldots f(x_n)$ is Gaussian.

A set of basis functions h transform the original feature vector x into a new feature vector h(x). A regression model based on Gaussian processes can be represented as follows:

$$Y = h(x)^T \beta + f(x),$$

where Y is the output vector, h(x) is a set of basis functions evaluated at all training points, β is the vector of basis function coefficients, f(x) is a zero mean Gaussian process with covariance function k(x, x′).

Then it is necessary to obtain the target distribution of the output vector. Based on the Gaussian process regression, an instance of response can be modeled as:

$$P(y_i \mid f(x_i), x_i) \sim N(y_i \mid h(x_i)^T \beta + f(x_i), \sigma^2),$$

where P is the posterior distribution, N is the normal distribution, matrix, σ^2 is error variance.

The Gaussian process regression is a probabilistic model. There is a latent variable $f(x_i)$ introduced for each observation x_i , that makes the model nonparametric. In vector form, this model can be represented as follows:

$$P(Y \mid f, X) \sim N(Y \mid H\beta + f, \sigma^2 I),$$

where Y, f, X, H are represented as follows:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}, f = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \cdot \\ \cdot \\ \cdot \\ f(x_n) \end{pmatrix}, X = \begin{pmatrix} x_1^T \\ x_1^T \\ \cdot \\ \cdot \\ \cdot \\ x_n^T \end{pmatrix}, H = \begin{pmatrix} h(x_1^T) \\ h(x_2^T) \\ \cdot \\ \cdot \\ \cdot \\ h(x_n^T) \end{pmatrix}$$

Therefore, to obtain a prediction by the studied model it is necessary to know the coefficients of the vector β, the error variance σ^2 and to be able to evaluate the covariance function (often this is a difficult task due to the so-called hyperparameters θ - unknown parameters that can vary). One of the methods for estimating the necessary parameters is to find the maximum of the following functional:

$$\hat{\beta}, \hat{\sigma}^2, \hat{\theta} = \underset{\beta,\sigma^2,\theta}{\arg\max} \log P(Y \mid X, \beta, \sigma^2, \theta),$$

where $\hat{\beta}, \hat{\sigma}^2, \hat{\theta}$ are the estimates of parameters, arg max is an argument of the maximum, log is a common logarithm.

Firstly, an estimate of the parameters β for the given values of σ^2 and θ is obtained by the formula:

$$\hat{\beta}(\sigma^2, \theta) = [H^T [K(X, X \mid \theta) + \sigma^2 I]^{-1} H]^{-1} \cdot H^T [K(X, X \mid \theta) + \sigma^2 I]^{-1} Y,$$

where the covariance function is written as K(X, X | θ) to explicitly indicate the dependence on θ and looks as follows:

$$K(X, X \mid \theta) = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & \ldots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \ldots & k(x_2,x_n) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ k(x_n,x_1) & k(x_n,x_2) & \ldots & k(x_n,x_n) \end{pmatrix}$$

Then, the functional presented above is maximized with respect to σ^2 and θ to obtain their estimates.

The Gaussian process regression is trained and tested on the same data division scheme as the two previous models under consideration. Based on the Gaussian process regression method, higher time costs for this model can be assumed.

The neural network model (multilayer perceptron) is a three-layer network with a sigmoidal activation function in the form of a hyperbolic tangent for a hidden layer and a linear activation function for the output layer, the Levenberg-Marquardt algorithm was used as a learning algorithm.

A detailed description of this model is given in [21], that addresses the issues of artificial neural networks to solve the task of natural gas composition analysis.

Recurrent neural networks (RNN) is a class of neural networks that can use their internal memory when processing input data [22]. The functioning of this class of neural networks is based on the use of previous network state to calculate the current one. A recurrent network can be considered as several copies of the same network, each of which transfers information to a subsequent copy. Currently, there are a large number of architectures of recurrent neural networks. Taking into account the computational difficulties encountered in developing this class of neural networks, it was proposed to consider a simple recurrent neural network first. The hidden elements have links directed back to the input layer in such type of network. This allows to take into account the previous state of the network during training. Mathematically, the process of saving information about the previous training step is as follows: at each i-th training step, the output value of the RNN hidden layer h_i is calculated taking into account the output value of the hidden layer in the previous step h_{i-1}:

$$h_i = f(W_h X_i + U_h h_{i-1} + b_{h0}),$$

where W_h, U_h, b_{h0} are parameters of the RNN hidden layer.

The output value at the i-th training step is calculated as follows:

$$y_i = W_{out} h_i + b_{out0},$$

where W_{out}, b_{out0} are parameters of RNN output layer.

The architecture of the considered RNS is shown in Fig. 1. The number of neurons at the input (n), hidden (k) and output (m) layers, the activation functions for the layers (for the hidden layer - sigmoidal function in the form of hyperbolic tangent, for the output layer – linear function), the learning algorithm (Levenberg-Marquardt) were chosen the same as for the neural network model in the form of a multilayer perceptron.

Taking into account the more complex RNN architecture in comparison with other models, the accuracy of the natural gas composition analysis is expected to increase.

A comparative analysis was proposed of a recurrent neural network with long short-term memory [23] to test the idea that increasing the complexity of the neural network architecture within one type of network (for example, RNN) does not lead to a significant improvement of the natural gas composition analysis. Long short-term memory (LSTM) is a special type of architecture of recurrent neural networks, that is capable of learning long-term dependencies. A more complex method is used to calculate both the output value of the hidden layer and the output value of the network as a whole in neural networks with a similar architecture. This method involves use of so-called gates. A gate is a special unit in LSTM architecture, that is implemented as a logistic function and operation of elementwise multiplication (Hadamard's product). The logistic function layer shows how much of the information coming from a particular unit should be transmitted further along the network. This layer returns values in the range from zero (information does not go further along the network structure at all) to one (information completely goes further along the network structure). There are three such gates in traditional LTSM architecture: a forget gate, an input gate and an output gate. The sigmoid function is often used as a logistic function for gates.

Let us take a closer look at the functioning of the LSTM unit. The input vector X_i, the long-term memory vector LTM_{i-1} (the state vector of the unit at the (i-1)-th step) and the vector of the working memory WM_{i-1} (the output vector of the unit at (i-1)-th step) come to LSTM unit at the i-th step of the model training. The forget gate and the input gate are used while calculating the long-term memory vector. Firstly, the forget gate is used to determine the proportion of long-term memory from the previous step, which should kept in use at

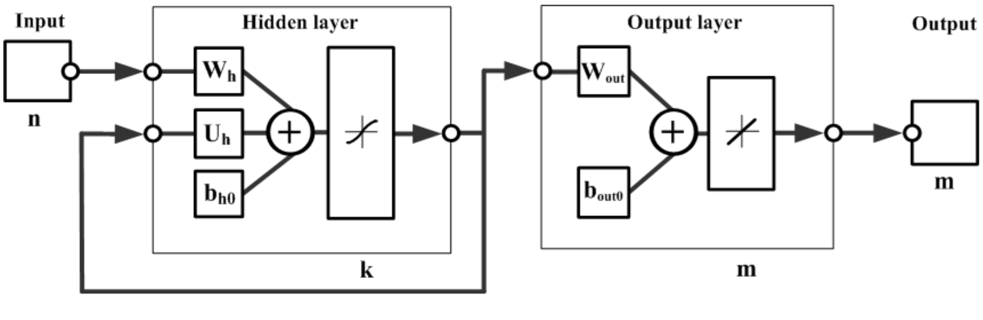

Fig. 1. Architecture of recurrent neural network (RNN) model

the current step. The forget gate is calculated by the formula:

$$forget_i = \sigma(W_f X_i + U_f WM_{i-1} + b_{f0}),$$

where σ is a sigmoid function of the forget gate, W_f, U_f, b_{f0} are parameters of the forget gate of LSTM unit.

After that, the proportion of information from the input data vector that can be added to long-term memory is determined:

$$LTM'_i = \tanh(W' X_i + U' WM_{i-1} + b'_0),$$

where tanh is an activation function in the form of hyperbolic tangent, W', U', b'_0 are LSTM unit parameters.

The input gate is calculated in order to estimate the useful proportion of the previous step that will be added to the long-term memory:

$$input_i = \sigma(W_{input} X_i + U_{input} WM_{i-1} + b_{input0}),$$

where σ is a sigmoid function of the input gate, W_{input}, U_{input}, b_{input0} are parameters of the input gate of LSTM unit.

Taking into account the performed operations, i.e. eliminating unnecessary information from the previous step and adding useful information from the current step, the vector of updated long-term memory can be calculated:

$$LTM_i = forget_i * LTM_{i-1} + input_i * LTM'_i,$$

where * is an elementwise multiplication operation.

After that, it is necessary to calculate the vector of working memory. An output gate is used for calculating the vector of working memory. It is necessary to calculate proportion of information from long-term memory that should be used at the current training step to calculate the vector of working memory. The output gate is calculated as follows:

$$output_i = \sigma(W_{output} X_i + U_{output} WM_{i-1} + b_{output0}),$$

where σ is a sigmoid function of output gate, W_{output}, U_{output}, $b_{output0}$ are parameters of output gate of LSTM unit.

Then, the vector of working memory is calculated at the current step:

$$WM_i = output_i * \tanh(LTM_i),$$

where tanh is an activation function in the form of hyperbolic tangent, * is an elementwise multiplication operation.

The calculated vectors of long-term memory LTM_i and working memory WM_i will go to the LTSM unit at the following training step.

The architecture of the LTSM unit is shown in Fig. 2. The general RNN architecture with long short-term memory is the same as for a simple RNN, taking into account an LSTM unit in the hidden layer.

The output value at the i-th training step for the RNN with the LTSM unit is calculated the same way as for a simple RNN:

$$y_i = W_{out} WM_i + b_{out0},$$

where W_{out}, b_{out0} are parameters of RNN output layer with LSTM unit.

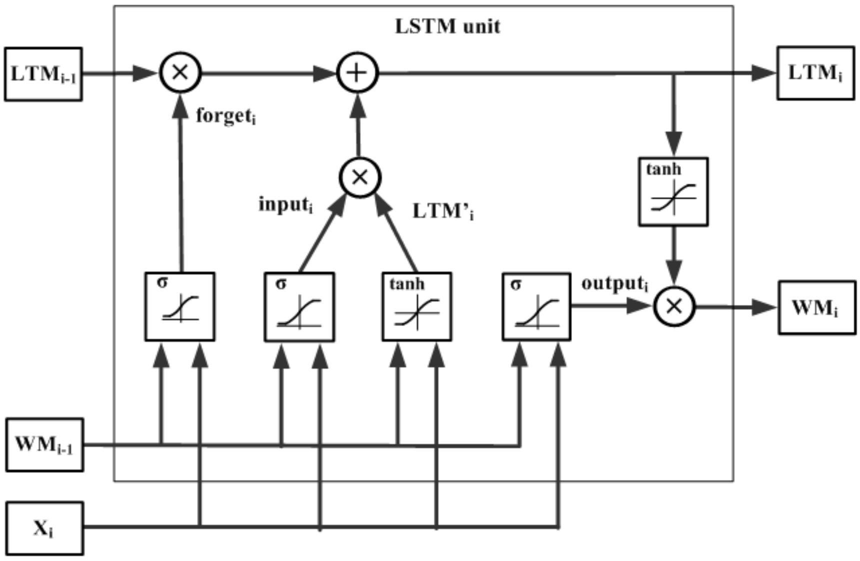

Fig. 2. Architecture of long short-term memory (LSTM) unit

Taking into account the more complex architecture of RNN with LTSM unit in comparison with a simple RNN, but with a similar principle of functioning of a neural network model, the time spent on training the model is expected to increase and accuracy characteristics of the natural gas composition analysis are expected to maintain.

3. Comparative analysis of the selected statistical models

All selected statistical models were trained on the same data generated according to the previously described requirements. The input parameters for the models were calculated using the NIST REFPROP software. The training of statistical models was carried out in the Matlab 2018a software. The selected models were trained on the same data several times, and then the average model training time and the average accuracy characteristics of the model for several training cycles were taken for increasing the analysis adequacy. A comparative analysis of the time spent on training for the studied models is presented in Table 3.

A comparative analysis of accuracy characteristics at the training stage for the models under study is shown in Table 4. Absolute errors (MAE, MaxAE) are given in units of determined concentrations (in %), relative errors (MAPE, MaxAPE) are given in %.

A comparative analysis of the accuracy characteristics at the testing stage for the studied models is shown in Table 5.

Table 3. Training time for the studied models

Studied model	Average training time
Multiparameter linear regression (LINREG)	6 seconds
Ridge regression (RIDGE)	4 seconds
Gaussian process regression (GPR)	45 minutes
Neural network model (multilayer perceptron) (ANN)	1,3 hours
Recurrent neural network (RNN)	3,5 hours
Recurrent neural network model with long short-term memory (LSTM)	5,6 hours

Table 4. Accuracy characteristics at the training stage for the studied models

Component	Characteristic	Model					
		LINREG	RIDGE	GPR	ANN	RNN	LSTM
Methane	MaxAE, %	5,383	5,373	0,581	0,491	0,184	0,184
	MAE, %	0,382	0,382	0,007	0,007	0,001	0,001
	MaxAPE, %	5,396	5,386	0,611	0,491	0,184	0,184
	MAPE, %	0,441	0,442	0,008	0,008	0,001	0,001
Nitrogen	MaxAE, %	6,464	6,450	0,362	0,247	0,253	0,241
	MAE, %	0,479	0,480	0,025	0,010	0,012	0,012
	MaxAPE, %	6,481	6,472	0,383	0,249	0,255	0,254
	MAPE, %	0,514	0,501	0,027	0,011	0,012	0,012
Propane	MaxAE, %	1,081	1,077	0,589	0,426	0,189	0,174
	MAE, %	0,107	0,107	0,011	0,007	0,005	0,005
	MaxAPE, %	1,095	1,087	0,589	0,446	0,183	0,171
	MAPE, %	0,109	0,109	0,011	0,009	0,004	0,004

Table 5. Accuracy characteristics at the testing stage for the studied models

Component	Characteristic	Model					
		LINREG	RIDGE	GPR	ANN	RNN	LSTM
Methane	MaxAE, %	5,531	5,399	0,592	0,511	0,361	0,295
	MAE, %	0,416	0,399	0,009	0,008	0,004	0,003
	MaxAPE, %	5,578	5,423	0,712	0,529	0,421	0,356
	MAPE, %	0,567	0,487	0,012	0,010	0,005	0,004
Nitrogen	MaxAE, %	6,691	6,689	0,353	0,236	0,241	0,233
	MAE, %	0,523	0,501	0,024	0,009	0,010	0,010
	MaxAPE, %	6,526	6,516	0,394	0,253	0,258	0,255
	MAPE, %	0,578	0,561	0,028	0,011	0,012	0,012
Propane	MaxAE, %	1,125	1,099	0,592	0,432	0,193	0,181
	MAE, %	0,109	0,109	0,012	0,007	0,005	0,004
	MaxAPE, %	1,239	1,102	0,596	0,448	0,188	0,175
	MAPE, %	0,131	0,115	0,011	0,009	0,004	0,004

Conclusion

The statistical models under study (multiparameter linear regression, ridge regression, Gaussian process regression, neural network model (multilayer perceptron), recurrent neural network, recurrent neural network with long short-term memory) were subjected to a comparative analysis to select the most suitable model for solving the task of natural gas composition analysis. Based on a comparison of the training time for the models under study, it was concluded that the model's training time increased while its architecture became more complicated.

Taking into account that in the proposed method for natural gas composition analysis [3], models train on theoretical data until the stage of applying the model to real data, increasing the training time within several hours does not impair the applicability of the method. Based on the comparison of the accuracy characteristics of the models both at the training stage and at the testing stage, it was concluded that the model error decreases with the complexity of its architecture. Summarizing the results of the comparative analysis of the statistical models, it is proposed to use a model with a more complex architecture (RNN) in working with real data. It is recommended to use neural network models with a similar architecture in similar tasks of the natural gas composition analysis, as well as in other tasks of determining the gas energy characteristics.

Further research is required in the field of developing a method for natural gas composition analysis using more complex neural network model for real data and adjusting the architecture parameters of the neural network model to solve the task of analyzing specific gas mixtures.

The author is grateful to the international cooperation projects BRISK II TA and Erasmus + 2017-1-SE01-KA107-034292 Staff Mobility between the universities for the opportunity to conduct this study.

References

1. GOST 31369-2008. Gaz prirodnyiy. Vyichislenie teplotyi sgoraniya, plotnosti, otnositelnoy plotnosti i chisla Vobbe na osnove komponentnogo sostava [Natural gas. Calculation of the calorific value, density, relative density and Wobbe number based on the composition.]. Moscow: Standartinform. 2008. 30 p.
2. Dörr H., Koturbash T., Kutcherov V. Review of impacts of gas qualities with regard to quality determination and energy metering of natural gas // Measurement Science and Technology. 2019. V. 30, №2. P. 1-20.
3. Koturbash T.T., Brokarev I.A. Metod opredeleniya svoystv i sostava prirodnogo gaza po izmereniyam ego fizicheskih parametrov [Method for determining the properties and composition of natural gas by measuring its physical parameters] // Datchiki i sistemyi [Sensors and systems]. 2018. № 6. C. 43-50.

4. Kostin V.N., Tishina N.A. Statisticheskie metodyi i modeli [Statistical methods and models]. Orenburg: GOU OGU, 2004. 138 p.
5. GOST P 8.662-2009 (ISO 20765-1:2005) Gaz prirodnyiy. Termodinamicheskie svoystva gazovoy fazyi. Metodyi raschetnogo opredeleniya dlya tseley transportirovaniya i raspredeleniya gaza na osnove fundamentalnogo uravneniya sostoyaniya AGA8 [Natural gas. Thermodynamic properties of the gas phase. Calculation methods for the transportation and distribution of gas based on the fundamental equation of state AGA8]. Moscow: Standartinform. 2010. 43 p.
6. Koturbash T.T., Brokarev I.A. Sravnitelnyiy analiz fizicheskih svoystv prirodnogo gaza i ekvivalentnyih emu psevdogazovyih smesey [Comparative analysis of the physical properties of natural gas and equivalent pseudogas mixtures] // Datchiki i sistemyi [Sensors and systems]. №3. 2019. P. 7-13.
7. Koturbash T., Bicz A., Bicz W. New instrument for measuring velocity of sound and quantitative characterization of binary gas mixtures composition // Measurement Automation Monitoring. 2016. P. 254-258.
8. Löfqvist T., Delsing J., Sokas K. Speed of sound measurements in gas — mixtures at varying composition using an ultrasonic gas flow meter with silicon based transducers // International Conference on Flow Measurement. Groningen, Netherlands. 2003.
9. Thermal Conductivity Gauge. Available at: http://www.xensor.nl (Accessed December 1, 2019).
10. Dynament Infrared Gas Sensors. Available at: https://www.dynament.com (Accessed December 1, 2019).
11. Bright Sensors BlueEye. Available at: https://www.brightsensors.com (Accessed December 1, 2019).
12. Mirzaei-Paiaman A., Salavati S. The application of artificial neural networks for the prediction of oil production flow rate // Energy Sources, Part A: Recovery, Utilization, and Environmental Effects. 2012. No. 34:19. P. 1834-1843.
13. Hribar R., Potočnik P., Šilc J., Papa G. A comparison of models for forecasting the residential natural gas demand of an urban area // Energy. 2018. Vol. 167. P. 511-522.
14. Vondráček J., Pelikán E., Konár O., Čermáková J., Eben, K., Malý, M., Brabec, M. A statistical model for the estimation of natural gas consumption // Applied Energy. 2008. No. 85(5). P. 362-370.
15. Aleardi, M. Analysis of different statistical models in probabilistic joint estimation of porosity and litho-fluid facies from acoustic impedance values // Geosciences. 2018. No. 8(11). P. 386-388.
16. Mitchell T. M. Machine Learning // McGraw-Hill Science/Engineering/Math. 1997.
17. Graybill F.A., Iyer H.K. Regression analysis // Concepts and applications, Duxbury Print. 1994.
18. Strizhov V.V., Kryimova E.A. Metodyi vyibora regressionnyih modeley [Regression model selection methods]. Moscow: VTS RAN, 2010. 60 p.
19. Hastie T., Tibshirani R., Friedman J. The Elements Of Statistical Learning: Data Mining, Inference and Prediction // Springer. 2009.
20. Rasmussen C. E., Williams C. K. Gaussian Processes for Machine Learning // The MIT Press. 2006.
21. Brokarev I.A. Iskusstvennyie neyronnyie seti dlya resheniya zadachi analiza komponentnogo sostava gazovyih smesey [Artificial neural networks for solving the problem of analyzing the composition of gas mixtures] // Upravlenie bolshimi sistemami [Large-scale Systems Control]. V. 80. Moscow: IPU RAN, 2019. P.98-115.
22. Callan R. The essence of neural networks (The essence of computing series) // Prentice Hall Europe. 1999.
23. Hochreiter S., Schmidhuber J. Long short-term memory // Neural computation. 1997. Vol. 9(8). P. 1735-1780.

Brokarev I. A. Assistant, PhD student of National University of Oil and Gas «Gubkin University», Moscow, 119991, Leninskiy prospect, 65, bl.1. Graduated from National University of Oil and Gas «Gubkin University» in 2017. 14 published articles. Topics of interest: information technologies, machine learning, natural gas analysis. E-mail: brokarev.i@gubkin.ru

Vaskovskii S. V. Senior research associate of V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, 117997, Profsoyuznaya street, 65. PhD. Graduated from Moscow Power Engineering Institute in 1986. 66 published articles. Topics of interest: information technologies, distributed computer systems, computer speech technologies. E-mail: v63v@yandex.ru