Gas Quality Determination Using Neural Network Model-based System

Ivan Brokarev[1] and Sergei Vaskovskii[2]

[1] National University of Oil and Gas "Gubkin University", Moscow, Russia
brokarev.i@gubkin.ru
[2] V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
v63v@yandex.ru

Abstract. The development of natural gas quality parameters determination system have been studied. The proposed system is based on neural network analysis and correlation analysis. The structure of the developed system is presented. The main blocks and subblocks of the proposed system are presented. The statistical model selection block and the architecture and parameters of the model selection block that are the most crucial stages of the proposed system are described in detail. The mathematical principles of the discussed statistical models are provided. The architectures of the studied neural network models are presented. The results of comparative analysis of different statistical models including neural networks are shown. The accuracy characteristics both for training and testing stages of the neural networks are calculated. The conclusion of final neural network architecture for the studied task was made. The results of testing of the proposed natural gas quality parameters determination system are provided. The steps for the further research of the discussed task are considered.

Keywords: Neural Network Analysis · Natural Gas Quality Analysis · Correlation Analysis · Statistical Model Selection

1 Introduction

The natural gas quality analysis is an important task for the gas industry. Slight fluctuations of natural gas composition and energy characteristics can lead to unexpected difficulties in calculating its cost indicators. Currently, a wide variety of different natural gas analysis systems are developed. Moreover, many alternative systems that are based on the correlation methods are under development [1]. The possibility to analyze gas quality in real time is the most significant benefit of this class of systems in comparison with systems based on the traditional gas chromatography methods. However, systems that are commonly used in gas industry have a number of drawbacks: expensive specialized equipment, significant amount of time of the analysis, the necessity of regular instrumentation calibration and checkout. Various statistical models are used in correlation methods because of high complexity of solving the task with traditional computational methods. The artificial neural networks are highly utilized in industrial and engineering applications [2,3]. The choice of statistical model for the gas quality determination is made by heuristic methods in most cases due to the lack of a general algorithm. That is why comparative analysis of statistical models for the discussed task is an urgent problem that should be solved for reaching the required goal.

This paper provides a structure and description of the main blocks of the proposed natural gas quality parameters determination system. The system is based on the method of determination of the properties and the composition of natural gas by measuring of its physical parameters [4]. The conclusions are drawn about further development of the proposed system.

2 Development of the Natural Gas Quality Parameters Determination System

The main structure of the proposed natural gas quality parameters determination system is shown in Fig. 1. The system consists of three blocks. We suggest using commercially available and relatively inexpensive sensors for natural gas physical parameters measurements to obtain necessary measurement data that are input data of the proposed system. The measurement data include following natural gas physical parameters: speed of sound, thermal conductivity and molar fraction of carbon dioxide. The aim of the system is to determine target natural gas quality parameters using input measurement data.

The first block (pseudogas composition determination) is the main block of the system that contains the majority of features of the proposed system. The task of this block are simplifying the studied object and minimizing amount of measured physical parameters and in its turn amount of applied sensors. This block we will describe below in more detail. The obtained equivalent pseudogas composition is transmitted to the next block where energy parameters calcula-

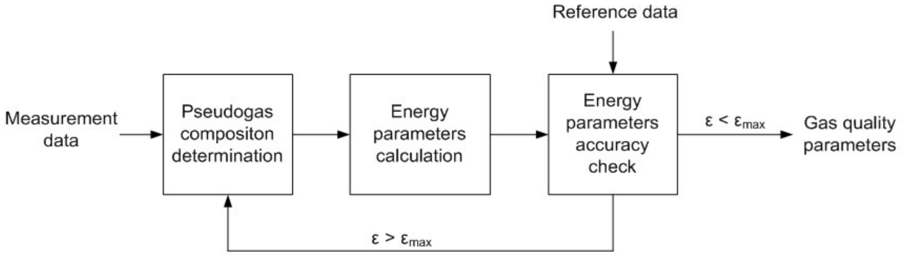
Reference data
Measurement data
Pseudogas compositon determination
Energy parameters calculation
Energy parameters accuracy check
ε < ε_max
Gas quality parameters
ε > ε_max

Fig. 1. Main structure of the proposed natural gas quality parameters determination system.

tion occurs. To calculate the energy parameters of the gas under study, NIST REFPROP software is used. The target energy parameters for the discussed task are volumetric superior calorific value and Wobbe index. These parameters along with partial gas composition and relative density are considered to be final gas quality parameters that system should determine. To calculate the quality parameters, the GERG-2008 gas state equation was used at standard temperature and pressure conditions. The amount of output parameters can be decreased to simplify the calculations or increased by adding volumetric inferior calorific value in special cases. The next step involves energy parameters accuracy check that occurs in the corresponding block. The calculated in previous block gas quality parameters are compared with reference data. Any data obtained from traditional natural gas analyzers, e.g. gas chromatographs, can be used as the reference data. The final error parameter ε is calculated to receive deviation of system parameters from reference parameters. That parameter is based on a number of accuracy characteristics including maximum absolute error (MaxAE), mean absolute error (MAE), maximum absolute percentage error (MaxAPE) and mean absolute percentage error (MAPE). In case of final error parameter is less than maximum limiting value ε_{max} the system provides the target gas quality parameters. In the opposite case, the stage of pseudogas composition determination is repeated. That includes a number of procedures that will be carried out until reaching the desired accuracy. The first block includes many subblocks that should be described separately. It's structure is shown in Fig. 2. The gas mixer block forms a natural gas composition. For the data formation a sample of gas mixtures based on the typical natural gas is simulated taking into account the permissible ranges of the molar fractions of the components by

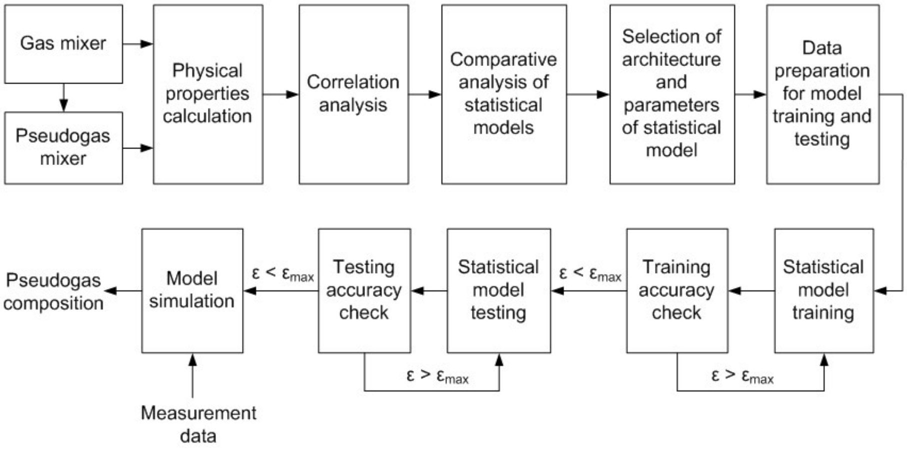
Gas mixer
Pseudogas mixer
Physical properties calculation
Correlation analysis
Comparative analysis of statistical models
Selection of architecture and parameters of statistical model
Data preparation for model training and testing
Statistical model training
Training accuracy check
ε > εmax
ε < εmax
Statistical model testing
Testing accuracy check
ε > εmax
ε < εmax
Model simulation
Measurement data
Pseudogas composition

Fig. 2. Structure of the block for pseudogas composition determination.

sorting out all possible combinations of components. Then the simulated gas mixtures are reduced to equivalent fourcomponent pseudogas mixtures in pseudogas mixer block. The physical properties calculation occurs in corresponding block. That process is similar to energy parameters calculation and includes calculation of theoretical values of parameters that will be used as measured. The correlation analysis is performed in corresponding block 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. The parameter list can be changed due to correlation analysis results. The next steps are comparative analysis of statistical models and model architecture and parameters selection. These two stages will be described below in detail. The data preparation stage that occurs in corresponding block includes data division on training, validation and test sets. It should be noted that prior to training the model, the data are cross-validated and normalized in order to be able to be used uniformly and improve the determination results of the statistical model. Moreover, the amount of initial data can be reduced due to desired ranges of gas components. The statistical model training stage involves selected model training using the selectable learning algorithm (Levenberg-Marquardt algorithm in default) on prepared at the previous stage data. The main tuning parameters of training are maximum number of training epochs (1000 in default), initial learning rate (0.001 in default), maximum validation failures (25 in default). The statistical model testing stage is the model simulation on the data that were not involved in training process. The accuracy check is provided both for training and testing stages. The procedure of accuracy check is similar to accuracy estimation in the energy parameters accuracy check block and involves calculation of error parameter. The final subblock of pseudogas composition determination block performs the function of model simulation on measurement data. The final parameter of the described subblock is the composition of equivalent pseudogas that will come to the next block of the proposed system.

3 Selection of Statistical Model, Its Architecture and Parameters

The stage of statistical model selection is considered to be the most crucial part of the proposed system that's why it will be described in detail. The choice of statistical 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. Therefore the methodology of comparative analysis should be developed. 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 criteria and characteristics to be used for model comparison, selection of statistical models for comparison. The data preparation stage is conducted in corresponding block as described above.

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 [5]. 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). 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). The next step of comparative analysis 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 based on an analysis of sources that address the problems arising when selecting statistical models for specific tasks of the gas industry, 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.

The multiparameter linear regression can be considered in the studied problem as a reference model. 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 $x_i, i = 1, \ldots, m$. The initial points are in an $m + 1$-dimensional space and are approximated by the m-dimensional hyperplane. The system of equations can be written in matrix form taking into account the error vector e:

$$Y = X\beta + e, \tag{1}$$

where β is a $(m+1)$-dimensional parameter vector, X is a matrix of row-vectors x_i.

To find an estimate of the regression 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. The estimation of the parameters of multi-parameter regression is the solution to this system using the least squares method, which may be shown as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{2}$$

The ridge regression is used in tasks with data redundancy as one of the methods of dimensionality reduction. 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 $X^T X$ 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 $X^T X$ matrix and obtain a more stable solution. The parameters of the regression model with regularization are found through minimizing the functional β^*:

$$\beta^* = argmin(||Y - X\beta||^2 + \tau||\beta||^2). \tag{3}$$

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

$$\beta^* = (X^T X + \tau I)^{-1} X^T Y, \tag{4}$$

where I is an identity matrix.

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.

The Gaussian process regression 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), $i = 1, \ldots, n$. The Gaussian process regression model explains the response by introducing latent variables, $f(x_i)$, $i = 1, \ldots, n$, 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 $m(x)$ and the covariance function $k(x, x')$ evaluated at x and x'. 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:

$$Y = h^T(x)\beta + f(x), \tag{5}$$

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. 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. 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 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:

$$\hat{\beta}, \hat{\sigma^2}, \hat{\theta} = \operatorname{argmax}(\log P(Y|X, \beta, \sigma^2, \theta)), \tag{6}$$

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

Firstly, an estimate of the parameters β for the given values of σ^2 and θ is obtained. Then, the functional presented above is maximized with respect to σ^2 and θ to obtain their estimates.

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. The neurons of each layer are connected with the neurons of the previous layer, and each input signal has a certain weight, that is the identical in this case for all input neurons due to the equal importance of all input values. Each neuron has an activation function, that argument is the input signal of the neuron. The chosen training algorithm is used to optimize the parameters of nonlinear regression models. The optimization criteria of the algorithm is the standard error of the model on the training set. The main idea of the algorithm is to achieve the desired local optimum by approximating the given initial parameter values.

The architecture of the used neural network model is shown in Fig. 3. The number of neurons in the input layer ($n = 3$) is chosen for the case when the concentration of carbon dioxide, the speed of sound and thermal conductivity

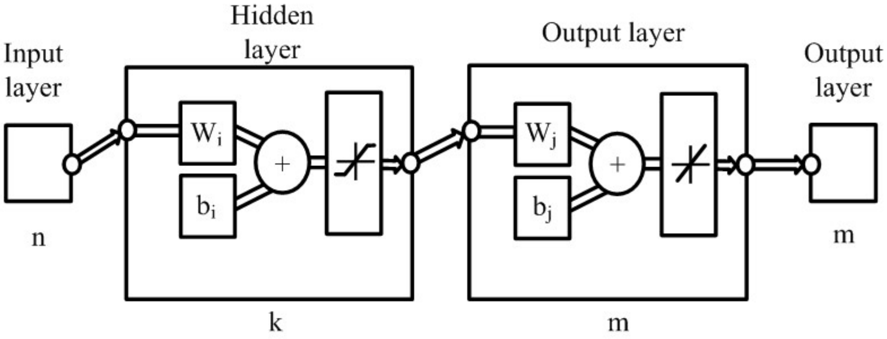

Fig. 3. Structure of the neural network model (multilayer perceptron).

are included in the input parameter vector. The number of neurons in the hidden layer ($k = 11$) is chosen for this particular model and is selected in accordance with the analysis of various models. The number of neurons in the output layer ($m = 3$) is chosen for the case of a four-component gas mixture. W_i and b_i are weights and bias factors for the hidden layer, W_j and b_j are for the output layer.

All selected statistical models were trained on the same data generated according to the previously described requirements. 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 1. A comparative analysis of accuracy characteristics at the training stage for the models under study is shown in Table 2. 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 3. The statistical models under consideration: multiparameter linear regression, ridge regression,

Table 1. Training time for the studied models.

Studied model	Average training time
Multiparameter linear regression (LINREG)	5 seconds
Ridge regression (RIDGE)	7 seconds
Gaussian process regression (GPR)	43 minutes
Neural network model (multilayer perceptron) (ANN)	1.2 hours

Table 2. Accuracy characteristics at the training stage for the studied models.

Component	Characteristic	Model			
		LINREG	RIDGE	GPR	ANN
Methane	MaxAE, %	5.383	5.373	0.581	0.491
	MAE, %	0.382	0.321	0.007	0.007
	MaxAPE, %	5.396	5.386	0.611	0.491
	MAPE, %	0.441	0.442	0.008	0.008
Nitrogen	MaxAE, %	6.464	6.450	0.362	0.247
	MAE, %	0.479	0.480	0.025	0.010
	MaxAPE, %	6.481	6.472	0.383	0.249
	MAPE, %	0.514	0.501	0.027	0.011
Propane	MaxAE, %	1.081	1.077	0.589	0.426
	MAE, %	0.107	0.107	0.011	0.007
	MaxAPE, %	1.095	1.087	0.589	0.446
	MAPE, %	0.109	0.109	0.011	0.009

Table 3. Accuracy characteristics at the testing stage for the studied models.

Component	Characteristic	Model			
		LINREG	RIDGE	GPR	ANN
Methane	MaxAE, %	5.531	5.399	0.592	0.511
	MAE, %	0.416	0.399	0.009	0.008
	MaxAPE, %	5.578	5.423	0.712	0.529
	MAPE, %	0.567	0.487	0.012	0.010
Nitrogen	MaxAE, %	6.691	6.689	0.353	0.236
	MAE, %	0.523	0.501	0.024	0.009
	MaxAPE, %	6.526	6.516	0.394	0.253
	MAPE, %	0.578	0.561	0.028	0.011
Propane	MaxAE, %	1.125	1.099	0.592	0.432
	MAE, %	0.109	0.109	0.012	0.007
	MaxAPE, %	1.239	1.102	0.596	0.448
	MAPE, %	0.131	0.115	0.011	0.009

regression based on Gaussian processes, a neural network model (multilayer perceptron) were put to comparative analysis for selection the most suitable model for solving the discussed task. It was concluded that the neural network model will be used as the main statistical model for solving the task.

The stage of architecture and parameters selection of statistical model was conducted similarly to previous procedure. The various applications of neural networks were analyzed taking into account the wide variety of architecture types of neural networks. For example, the convolutional neural networks were excluded from consideration due to the application of this type of neural networks mainly for recognition and classification tasks. In addition to the above-described neural network architecture in the form of a multilayer perceptron, a simple recurrent neural network and a recurrent neural network with long short-term memory were chosen for analysis.

Recurrent neural networks (RNN) is a class of neural networks that can use their internal memory when processing input data. 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}:

$$h_i = f(W_h X_i + U_h h_{i-1} + b_{h0}), \qquad (7)$$

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:

$$y_i = W_{out} h_i + b_{out0}, \tag{8}$$

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

The architecture of the considered RNN is shown in Fig. 4. 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. A comparative analysis was proposed of a recurrent neural network with long short-term memory 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.

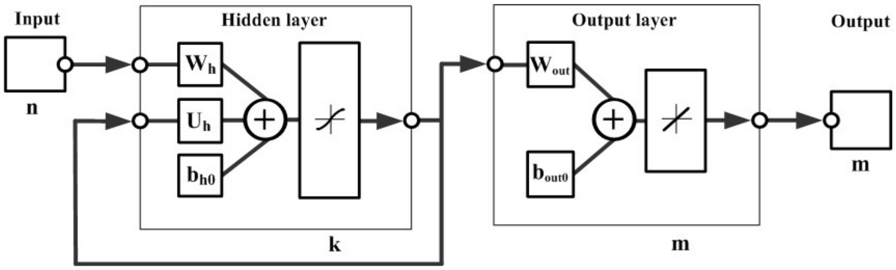

Fig. 4. Architecture of recurrent neural network model.

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

$$forget_i = \sigma(W_f X_i + U_f WM_{i-1} + b_{f0}), \tag{9}$$

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.

$$LTM'_i = \tanh(W' X_i + U' WM_{i-1} + b'_0), \tag{10}$$

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. The formula for the input gate is similar to the forget gate taking into account sigmoid function and parameters of the input gate. 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:

$$LTM_i = forget_i * LTM_{i-1} + input_i * LTM'_i, \tag{11}$$

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 similarly to forget and input gate taking into account sigmoid function and parameters of the output gate. Then, the vector of working memory is calculated at the current step:

$$WM_i = output_i * \tanh(LTM_i). \tag{12}$$

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. 5. 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.

The comparative analysis of different architectures was carried out similarly to comparative analysis for the selection of a statistical model. A comparative

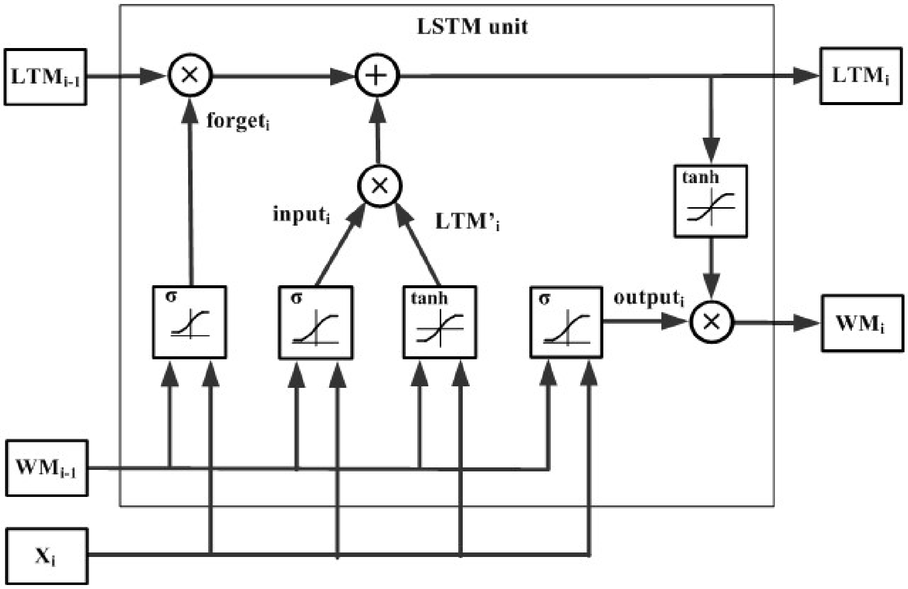

Fig. 5. Architecture of long short-term memory unit.

Table 4. Training time for the studied model architectures.

Studied model	Average training time
Neural network model (multilayer perceptron) (ANN)	1.2 hours
Recurrent neural network (RNN)	3.5 hours
RNN model with long short-term memory (LSTM)	5.6 hours

Table 5. Accuracy characteristics at the training stage for the studied model architectures.

Component	Characteristic	Model		
		ANN	RNN	LSTM
Methane	MaxAE, %	0.491	0.184	0.184
	MAE, %	0.007	0.001	0.001
	MaxAPE, %	0.495	0.183	0.182
	MAPE, %	0.008	0.001	0.001
Nitrogen	MaxAE, %	0.247	0.253	0.241
	MAE, %	0.010	0.012	0.011
	MaxAPE, %	0.249	0.255	0.254
	MAPE, %	0.011	0.012	0.011
Propane	MaxAE, %	0.426	0.189	0.174
	MAE, %	0.007	0.005	0.004
	MaxAPE, %	0.446	0.183	0.171
	MAPE, %	0.009	0.004	0.004

analysis of the time spent on training for the studied models is shown in Table 4. A comparative analysis of accuracy characteristics at the training stage for the model architectures under study is shown in Table 5. A comparative analysis of the accuracy characteristics at the testing stage for the studied model architectures is shown in Table 6. The statistical models under consideration:

Table 6. Accuracy characteristics at the testing stage for the studied model architectures.

Component	Characteristic	Model		
		ANN	RNN	LSTM
Methane	MaxAE, %	0.511	0.361	0.191
	MAE, %	0.008	0.004	0.001
	MaxAPE, %	0.529	0.421	0.216
	MAPE, %	0.010	0.005	0.002
Nitrogen	MaxAE, %	0.236	0.241	0.233
	MAE, %	0.009	0.010	0.010
	MaxAPE, %	0.253	0.258	0.255
	MAPE, %	0.011	0.012	0.011
Propane	MaxAE, %	0.432	0.193	0.181
	MAE, %	0.007	0.005	0.004
	MaxAPE, %	0.448	0.188	0.175
	MAPE, %	0.009	0.004	0.004

neural network model (multilayer perceptron), recurrent neural network, recurrent neural network model with long short-term memory were put to comparative analysis for selection the most suitable model for solving the discussed task. It was concluded that the simple recurrent neural network model will be used as the main statistical model for solving the task.

In this paper we considered models that architecture has one hidden layer. The dimension of the input layer is determined by the number of input physical parameters. The dimension of the output layer is determined by the number of determined concentrations of gas components. Therefore the number of neurons in the hidden layer was used as the main tunable parameter for the model. The accuracy characteristics are given in Table 7 only for the testing stage due to the slight difference in training time and accuracy at the testing stage. It was determined that the model with eleven neurons in the hidden layer has the highest accuracy based on the results of the study.

To sum up, the simple recurrent neural network with three layers and eleven neurons in hidden layer was chosen as the main statistical model.

4 Testing of the Proposed Natural Gas Quality Parameters Determination System

The proposed natural gas quality parameters determination system design and testing were carried out in the Matlab 2019b software with NIST REFPROP

Table 7. Accuracy characteristics at the testing stage for the different number of neurons in hidden layer.

Component	Characteristic	Number of neurons		
		5	11	25
Methane	MaxAE, %	0.467	0.361	0.499
	MAE, %	0.006	0.004	0.008
	MaxAPE, %	0.481	0.421	0.511
	MAPE, %	0.009	0.005	0.010
Nitrogen	MaxAE, %	0.179	0.241	0.223
	MAE, %	0.008	0.010	0.009
	MaxAPE, %	0.193	0.258	0.261
	MAPE, %	0.010	0.012	0.012
Propane	MaxAE, %	0.318	0.193	0.421
	MAE, %	0.006	0.005	0.007
	MaxAPE, %	0.371	0.188	0.439
	MAPE, %	0.008	0.004	0.009

plug-in. The main aim of testing is to verify system efficiency on the theoretical data. The initial data include 137214 gas mixtures that are based on typical natural gas. The ranges of its components are the following: 90-100% for methane, 0-3% for nitrogen and ethane, 0-1% for carbon dioxide and propane, 0-0.5% for butane and pentane, 0-0.2% for hexane. These mixtures were transformed to fourcomponent pseudogas mixtures. Then physical parameters of both types of mixtures were calculated. The number of parameters exceeds the number of statistical model input parameters to verify the previous results of correlation analysis. Additional physical parameters are dielectric permittivity, dynamic viscosity and isobaric heat capacity. The conducted correlation analysis proved the results of previous research. Speed of sound, thermal conductivity and molar fraction of carbon dioxide were selected as input parameters for the next stages.

The simple recurrent neural network with default architecture and parameters was chosen as working model. On the next step, the initial data was reduced to 111000 gas mixtures by eliminating gas mixtures with composition not close to the natural gas, e.g. pure methane. Then the data was divided on two sets for training and testing. The special data set was formed for simulation stage. It included 200 gas mixtures with calculated physical parameters. The selected recurrent neural network was trained, tested and simulated on the corresponding sets with accuracy characteristics shown in table 8. Each procedure was started only when the previous procedure (training in case of testing and testing in case of simulation) was successful. Carbon dioxide errors were set to zero, because the content of this component is input value and considered to be known. The calculated composition of simulation set was transmitted to energy parameters calculation block. Theoretical values of natural gas energy parameters was used as reference data. The volumetric superior calorific value and Wobbe index were calculated using determined pseudogas composition and compared with reference data. The accuracy of determination of target gas quality parame-

Table 8. Model accuracy characteristics at the training, testing and simulation stages.

Component	Characteristic	Stage		
		Training	Testing	Simulation
Methane	MaxAE, %	0.423	0.496	0.581
	MAE, %	0.007	0.008	0.012
	MaxAPE, %	0.531	0.625	0.751
	MAPE, %	0.008	0.010	0.014
Nitrogen	MaxAE, %	0.286	0.374	0.517
	MAE, %	0.011	0.012	0.018
	MaxAPE, %	0.301	0.372	0.521
	MAPE, %	0.013	0.015	0.022
Propane	MaxAE, %	0.251	0.333	0.491
	MAE, %	0.007	0.009	0.014
	MaxAPE, %	0.237	0.299	0.456
	MAPE, %	0.006	0.008	0.014

ters (deviation between determined by system and reference values) is shown in Fig.6 (for volumetric superior calorific value) and in Fig. 7 (for Wobbe index). The maximum absolute error of gas quality parameters determination (0.0364 MJ/m3 for calorific value and 0.0914 MJ/m3 for Wobbe index) is less than the allowable error that is equal to 0.1 MJ/m3. The allowable error is permissible deviation of gas quality parameters determination for the first accuracy class according to current regulatory document.

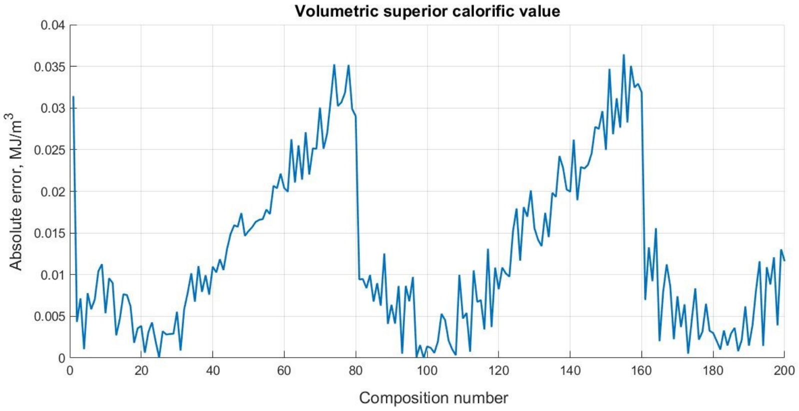

Fig. 6. The accuracy of determination of volumetric superior calorific value by proposed system.

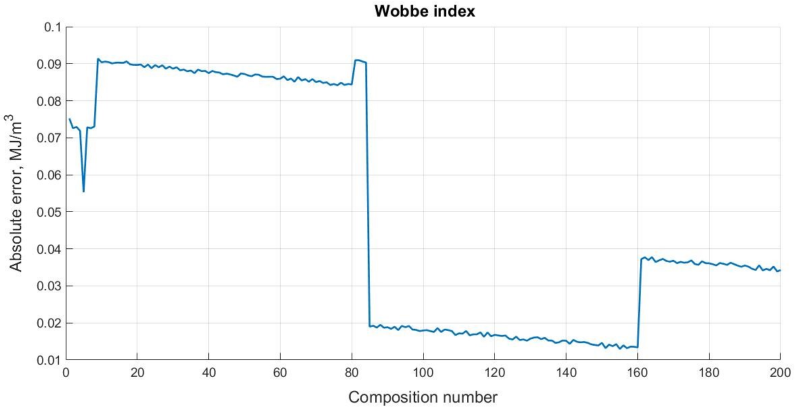

Fig. 7. The accuracy of determination of Wobbe index by proposed system.

5 Conclusion

The natural gas quality parameters determination system was presented. The main advantages of the proposed system in comparison with the commonly used gas quality determination methods are the high adaptability and capability of operation in real time. The target gas quality properties, volumetric superior calorific value and Wobbe index determined by the system were compared with reference data. The system showed acceptable performance on theoretical data. Further research is required in the field of testing the model on experimental data and adjusting system algorithms to solve the task of analyzing specific gas mixtures.

The authors are grateful to BRISK II TA and Erasmus + 2017-1-SE01-KA107-034292 Staff Mobility for providing the opportunity to conduct this study.

References

1. Dorr 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 **30**(2), 1–20 (2019)
2. Suzuki K.: Artificial Neural Networks - Industrial and Control Engineering Applications. InTech, USA (2011)
3. Cranganu C., Breaban M., Luchian H.: Artificial Intelligent Approaches in Petroleum Geosciences. Springer, Switzerland (2015)
4. Koturbash T.T., Brokarev I.A.: Method for determining the properties and composition of natural gas by measuring its physical parameters. Sensors and systems **6**, 43–50 (2018)
5. Mitchell T. M.: Machine Learning. McGraw-Hill Science/Engineering/Math, USA (1997)