Adv Syst Sci Appl 2019; **04**; 14-24
Published online at http://ijassa.ipu.ru/index.php/ijassa/article/view/754

Distributed data gathering system to analyze natural gas composition

Ivan Brokarev[1*], Sergei Vaskovskii[2]

[1)] *National University of Oil and Gas «Gubkin University», Moscow, Russia*

E-mail: brokarev.i@gubkin.ru

[2)] *V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia*

E-mail: v63v@yandex.ru

Received June 07, 2019; Revised September 30, 2019; Published December 31, 2019

Abstract: Problems of development of gas analysis system based on computing fabrics have been studied. The structure of the data gathering system that will be used for the natural gas composition analysis is proposed. The most significant features and the main advantages of the proposed system are described. The most appropriate statistical models that can be used in solving the task of natural gas composition analysis are presented. The consecutive stages for statistical model development and correlation analysis results are shown. Algorithms have been developed for effective solution for a number of gas analysis tasks for distributed systems. The mathematical basis of the learning algorithm and the architecture of the proposed neural network model are described. The different training completion cases for the developed neural network model are shown. Recommendations are made for effective application of neural network tools as the main models for solving the task of the natural gas composition analysis. Opportunities for further development of the proposed system are considered.

Keywords: Distributed systems, data gathering system, composition analysis, neural network analysis.

1. INTRODUCTION

Currently a variety of different gas analysis systems are under development [2]. Distributed small size computerized systems are widely used to develop automatic process control systems including gas analysis systems. Real time systems are the most effective systems to analyze gas. The Modcon Systems wireless gas analysis system is one of them; the system monitors the technological process online [7].

However, in spite of the past gains in this field, it is time-consuming to develop distributed real-time control and gas analysis systems. Existing systems have a number of disadvantages: it is necessary to use expensive specialized equipment, the system is oriented mainly to monitor the technological process, the system detects limited amount of gas components. So, it is worthwhile to develop distributed gas analysis systems to overcome these shortcomings. It is one of the most complicated task to develop a distributed hardware and software for distributed microcomputerized systems. The main task solved by a distributed data gathering system is to provide access to a large amount of data in real time.

This paper describes difficulties the engineers face while developing a distributed data gathering system to analyze natural gas composition; we take into account the specific features of such systems and how to apply neural network tools to investigate the features.

[*] Corresponding author: Ivan Brokarev (brokarev.i@gubkin.ru)

2. DEVELOPMENT OF THE DISTRIBUTED DATA GATHERING SYSTEM FOR NATURAL GAS COMPOSITION

Let us see how to develop a distributed data gathering system to analyze natural gas composition. We want the system to be automated and based on the proposed data gathering system. The automated control system can be both information-logical system to monitor the technological process of gas transport and information-computing system to calculate a number of qualitative natural gas parameters, for example, the gas calorific value. We suggest to use a multiplex data gathering system for the task. In particular, multiplex systems, that utilize individual signal processing techniques at each measuring channel, are currently the most widespread systems.

The most important features of distributed systems are the following:

- transparency of the system – ability to hide physical distribution of resources, errors concerning access to resources and operation itself, duplication of resources, difficulties concerning concurrent operation of several users with a single resource;
- scalability of the system – dependence of system characteristics on the number of its users and involved resources;
- system security – data integrity protection, proper fall-over protection and error recovery capability;
- openness of the system – availability of full description of interfaces and services of system operation;
- system real-time operation – ability to respond to unpredictable flow of external events within predictable time.

It should be noted that such features are inherent for the majority of distributed systems that can be applied to analyze gas. The main differences of the proposed system from the existing systems are commercially available and relatively inexpensive sensors as hardware components. In addition, our system is able to compute in real-time a number of natural gas energy characteristics, including its calorific value. Finally, our system can detect main natural gas components (for example, eight components should be detected for Russian natural gas [1]). All these can be detected and computed in real-time.

When developing the data gathering system the process is considered to be carried out in several stages. The main stage is dedicated to independent development and debugging of interconnected components of applied software and hardware.

The software of the designed data gathering system consists of the following main components:

- programming of the low level of data gathering (sensor level);
- interaction of the low and high levels (sensor level and data analysis level);
- data analysis after monitoring, data documenting and data archiving.

The structure of the proposed distributed system is shown in fig. 2.1. Two measurement chambers with connected sensors are located at the first level of our system. Two chambers are necessary to verify the collected measurement data. An important subsystem of our data gathering system is the sensor system. We suggest to use commercially available and relatively inexpensive sensors to measure physical parameters of gas mixtures – speed of sound [4], thermal conductivity [11], and molar fraction of carbon dioxide [3]. Our data collecting system is located at the second level of the proposed system. The main functions of that subsystem are to scan for the information and transit the information to the higher level. The system processes the gathered data at the third level. To process the data is to eliminate errors from the measuring results, to average the collected measurement data, to compute the accessory parameters and visualize the collected data. The algorithms to analyze and evaluate the collected data are implemented at the fourth level of the system. These algorithms will be described later. At that level we compute how accurate the designed neural network is. This accuracy output shows if our model is adequate.

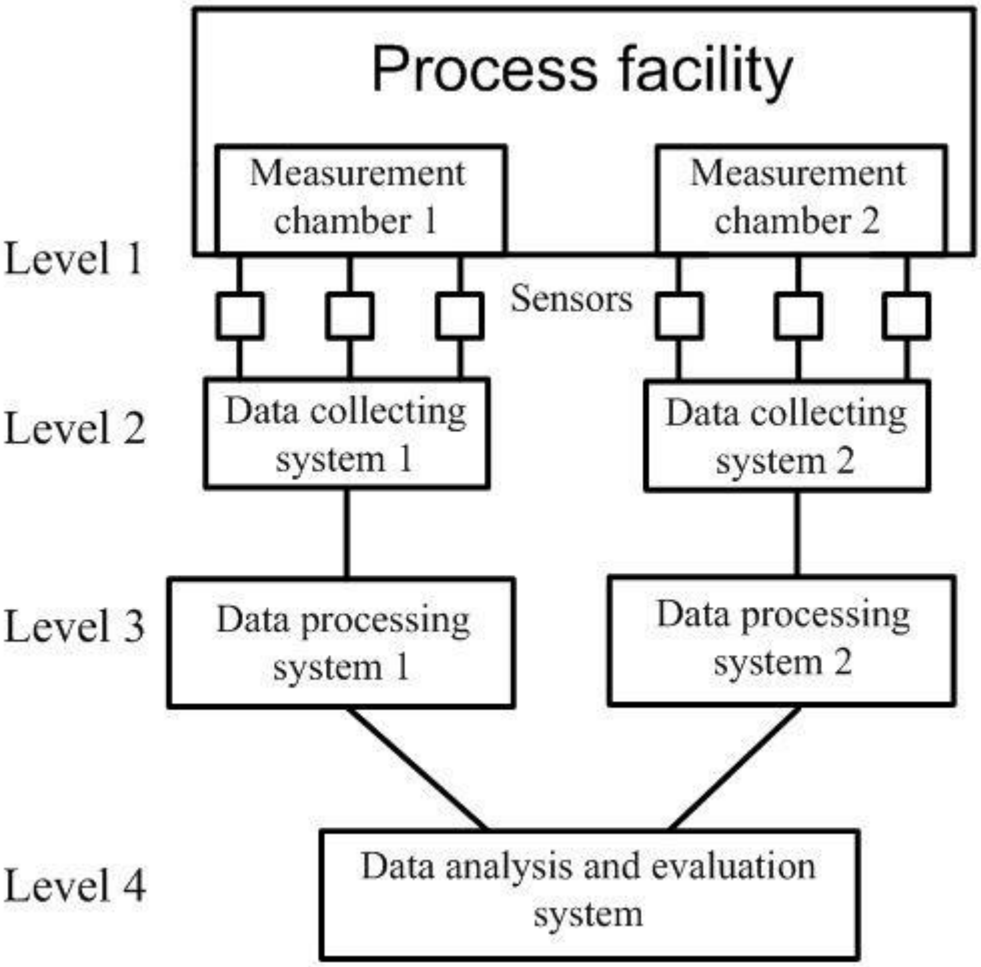

Fig. 2.1. Structure of the proposed distributed system

3. NEURAL NETWORK TOOLS FOR NATURAL GAS COMPOSITION ANALYSIS

Currently, artificial intelligence solves a variety of applied problems. One of the important areas of artificial intelligence are statistical models. Among a large variety of these models, we distinguish artificial neural networks (ANN) - mathematical models built on the principles of how biological neural networks are organized and function. ANN are trained to detect complex dependencies between input and output. This ability is one of the main advantages of neural networks over traditional algorithms.

One of the promising tasks, where ANNs are applied, is to analyze component composition of natural gas. If gases from different fields and sources are mixed during transportation and storage, the natural gas composition changes from that on the operation sites. This makes it harder to account the natural gas.

Since gas composition determines gas quality indicators for the equipment to operate properly and optimally, we need to measure or calculate certain properties of natural gas to transport the natural gas via a pipeline. This makes it very important to determine component composition of gas in real time; to solve this problem, correlation methods are currently being developed to analyze gas quality [2, 5]. Statistical models, in particular ANN, are often used to determine the desired properties and composition of natural gas by measuring the physical parameters of the gas.

The fourth level of the shown distributed system carries out algorithms for determination of the target composition by measuring input physical parameters of natural gas. A statistical method is proposed as the main method for determination of the target composition. Simplicity of mathematical calculations is the main advantage of the statistical method over traditional algorithms. The main disadvantage - the need for a large amount of initial data - is eliminated by conducting the measurements via two channels.

The following models can be used as statistical models for solving the task under discussion:

- multi-parameter linear regression can serve as a reference model. Its results can be used to compare accuracy of other regression models. For this model, it is necessary to calculate each component of the gas mixture separately. This is inefficient in comparison to other models;
- model based on the support vector machine (SVM). This is a set of similar supervised machine learning algorithms used to classify and regress analysis tasks. The main idea of the method is to translate source vectors into a higher dimension space and to search separating hyperplanes with a maximum gap in this space. We don't choose this model because of its significant shortcomings, in particular, complex interpretation of model parameters; also the model is applicable only to solve problems with two classes;
- gaussian process regression (GPR);
- ridge regression;
- neural network model.

On the basis of the previous research [5], it can be concluded that neural network is most effective as the main statistical model in the discussed task because of its learning capability and scalability in comparison with other models.

To develop a model to analyze gas mixture composition, we implement these consecutive stages:

- select data to train the model;
- choose model architecture;
- select a model training method;
- assess accuracy of the model.

The first stage consists in selecting input and output data for the model. We select a physical parameter as input, if there is a relationship between this gas parameter and the gas composition, as well as if the parameter can be measured with commercially available and relatively inexpensive measurement devices [3,4,11]. To determine if the parameters and the composition of the natural gas are related, a correlation analysis is carried out to verify inter dependencies within the sets of the input and output parameters and if the parameters are connected with each other.

All data is cross-validated before it is used to train the model. Cross-validation is a method to evaluate an analytical model and its behavior on independent data. This procedure is as follows: available data is divided into k parts, then the model is trained on the k - 1 parts of the data, and the rest of the data is used to test the model. The procedure is repeated k times; each of the k pieces of the data is used to test the model. As a result, we assess effectiveness of the model if the available data is used uniformly. For this task, the sample of N = 700000 elements was divided into k = 10 parts. Moreover, the training sample was normalized before training to improve results of prediction of the designed neural network model. The ranges of gas mixture components for the training samples are shown in Table 3.1.

Table 3.1. Ranges of components molar fractions for training sample

Component	Molar fraction, %
Methane	70 – 100
Ethane	0 – 10
Propane	0 – 5
Carbon dioxide	0 – 10
Nitrogen	0 – 10

The aim is to eliminate possible multicollinearity of the parameters - linear interrelation of two or several variables. Multicollinearity can lead to undesirable consequences, since the parameter estimates become unreliable. This means that the standard error increases, and it becomes impossible to isolate how the factors influence the effective indicators. We use correlation analysis (see Table 3.2) with carbon dioxide molar fraction, sound velocity, and thermal conductivity coefficient as the input parameters. The output parameters are molar fractions of the gas mixture components.

Table 3.2. The correlation analysis results

	Speed of sound, m/s	Thermal conductivity, mW/(m*K)	Methane molar fraction, %	Nitrogen molar fraction, %	Carbon dioxide molar fraction, %	Ethane molar fraction, %	Propane molar fraction, %
Speed of sound, m/s	1	0,99	0,93	-0,24	-0,61	-0,33	-0,67
Thermal conductivity, mW/(m*K)	0,99	1	0,91	-0,15	-0,53	-0,45	-0,71
Methane molar fraction, %	0,93	0,91	1	-0,52	-0,52	-0,48	-0,49
Nitrogen molar fraction, %	-0,24	-0,15	-0,52	1	0	0	0
Carbon dioxide molar fraction, %	-0,61	-0,53	-0,52	0	1	0	0
Ethane molar fraction, %	-0,33	-0,45	-0,48	0	0	1	0
Propane molar fraction, %	-0,68	-0,75	-0,49	0	0	0	1

The results of correlation analysis are shown on fig. 3.2 for parameters that have not been chosen as input parameters because of low correlation (speed of sound is shown as reference parameter for comparison). The results of correlation analysis are shown on fig. 3.3 for parameters that have been chosen as input parameters because of high correlation (methane concentration is shown as reference parameter for comparison).

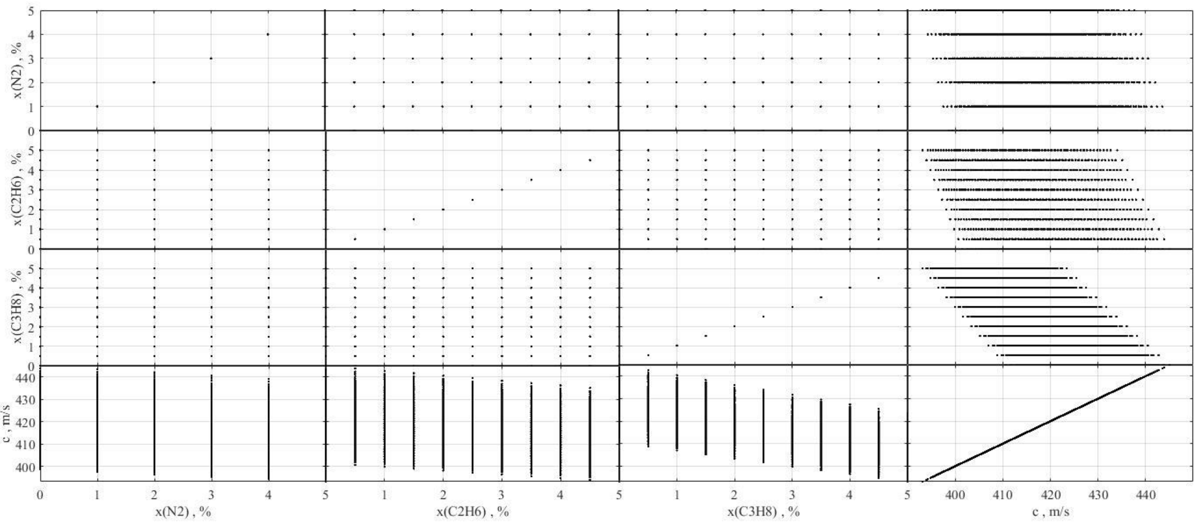

Fig. 3.2. The correlation analysis results for unchosen as input parameters

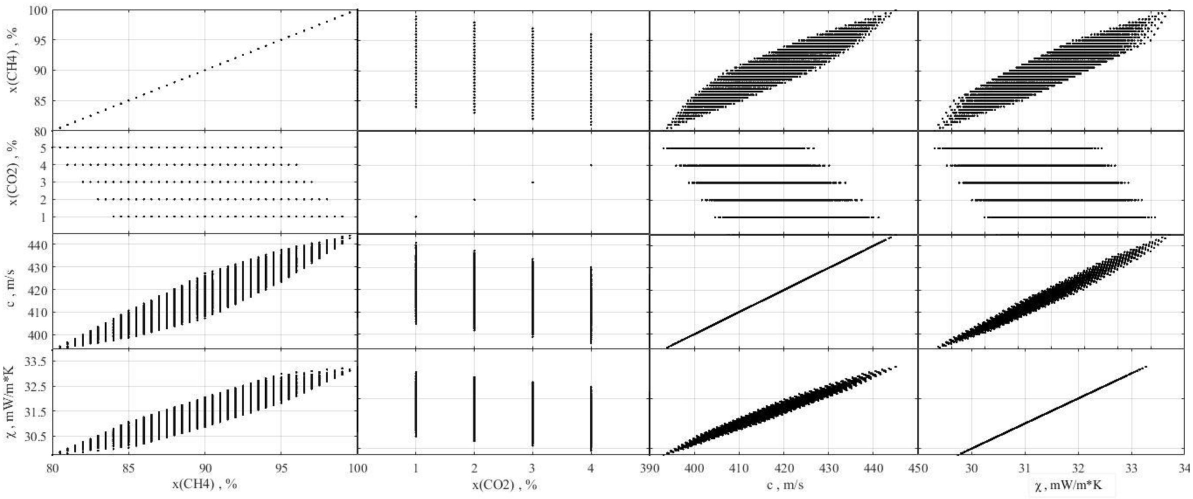

Fig. 3.3. The correlation analysis results for chosen as input parameters

This neural network model is proposed as the main statistical model in the task [9]. To solve the problem, a three-layer network (multilayer perceptron) was chosen, which includes two neurons in the input layer, where two is the number of components of the input vector, eleven neurons in the second layer and four neurons in the output layer, where four is the number of components of the output vector. The input layer is a vector X = [x0 x1 x2]. The neurons of each next layer are connected to the neurons of the previous layer, and each input signal has a certain weight. The weight is the same for all input neurons since all input variables are equally important. Each neuron has an activation function; the function argument is the input signal of the neuron. For the neurons of the hidden layer, the activation function is sigmoidal in the form of a hyperbolic tangent. For the neurons of the output layer, the activation function is linear. For training the network the Levenberg-Marquardt algorithm was chosen.

This algorithm optimizes parameters of nonlinear regression models. Note that in this algorithm, the optimization criterion is the root-mean-square error of the model on a training sample. The basic idea of the algorithm is as follows: to minimize the error locally, the initial values of the parameters are approximated.

Let a regression sample be a set of pairs of an independent variable X and a dependent variable Y, and the regression model is a continuously differentiable function F. It is necessary to find the value of the parameter vector W, where the error function Fε reaches its local minimum:

$$F_\varepsilon = \sum_{i=1}^{N} (Y_i - F(X_i, W) \tag{3.1}$$

On the first iteration of the algorithm, the initial vector of parameters W_0 is specified. On each following iteration, the vector is replaced by the vector W_0 + ΔW. To estimate the increment ΔW, the following approximation of F is used:

$$F(W_0 + \Delta W, X) - F(W_0, X) = J * \Delta W \tag{3.2}$$

where J is the Jacobian of F.

The increment ΔW at the minimum of Fε is zero. So, to find the subsequent value of the increment ΔW, it is necessary to set the vector of partial derivatives of Fε over W to zero. Then we differentiate this expression over W and set the partial derivative to zero. After all transformations, we get the expression for ΔW:

$$\Delta W = (J^T * J)^{-1} * J^T * (Y - F(W)) \tag{3.3}$$

For this algorithm, the condition number of the matrix is important; the number shows how close the matrix is to the partial rank matrix (for square matrices it shows how close the matrix is to degeneracy). Since the condition number of the matrix J^T*J is equal to the squared condition number of the matrix J, the matrix J^T*J may turn out to be degenerate. For this reason, in this algorithm we introduced a regularization parameter λ; the parameter is greater than or equal to zero. This parameter is selected on each iteration of the algorithm. Given the regularization parameter, expression for ΔW takes the following form:

$$\Delta W = (J^T * J + \lambda * E)^{-1} * J^T * (Y - F(W)) \tag{3.4}$$

where E is a unity matrix.

There is a modification of this method, where the regularization parameter is multiplied by the matrix D, a diagonal matrix with the elements that coincide with the diagonal elements of the matrix J^T*J. This approach is used to reduce the effect of the regularization parameter on the ΔW value.

It is necessary, however, to note that this algorithm converges slower if the step is constant, which is a disadvantage. The problem is solved by introducing a coefficient K; the coefficient determines the step length and makes the method converge faster. Given the two amendments described, the expression for ΔW takes the following form:

$$\Delta W = K * (J^T * J + \lambda * D)^{-1} * J^T * (Y - F(W)) \tag{3.5}$$

The value of the vector W at the last iteration of the algorithm is the target. It is reached either if the calculated increment ΔW is less than the specified value, or if the error function Fε is less than the specified value for the vector W.

The architecture of the neural network model is shown in fig. 3.4. The number of neurons in the input layer is n = 3 for the case when the vector of input parameters contains carbon dioxide molar fraction, sound velocity, and thermal conductivity. The number of neurons in the hidden layer of our model is 11; to get this number, many different models were analyzed. The number of neurons in the output layer is m = 4 for the case of a five-component gas mixture.

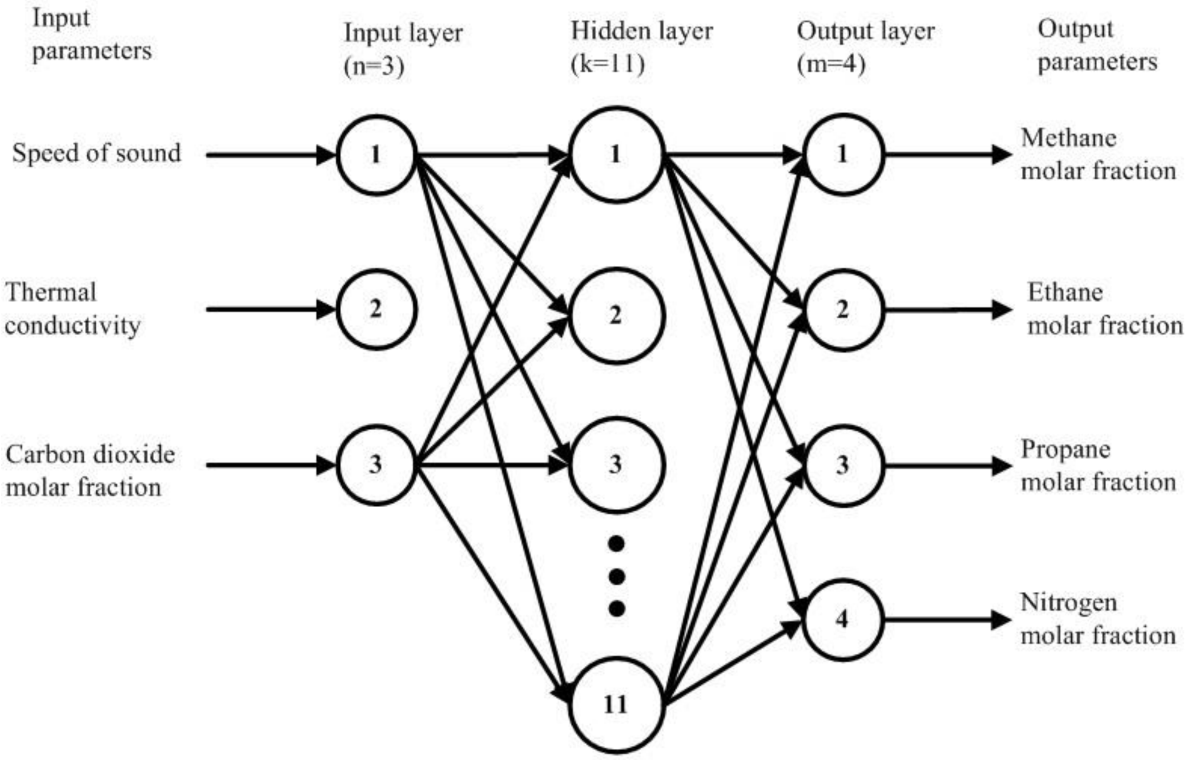

Fig. 3.4. The architecture of the proposed neural network model

The next step is to train the network on the previously obtained data using the Levenberg-Marquardt algorithm. Before that it is necessary to set the stopping criteria; for example, the criteria can be the maximum deviation value, when training is considered complete. In our case, to estimate the value, the sum of the squared deviations of the network outputs from the true values is calculated. Also, when training the model, the maximum number of training cycles may be selected as the stopping criteria. However, as can be seen from fig. 3.5, this choice makes the network longer to train.

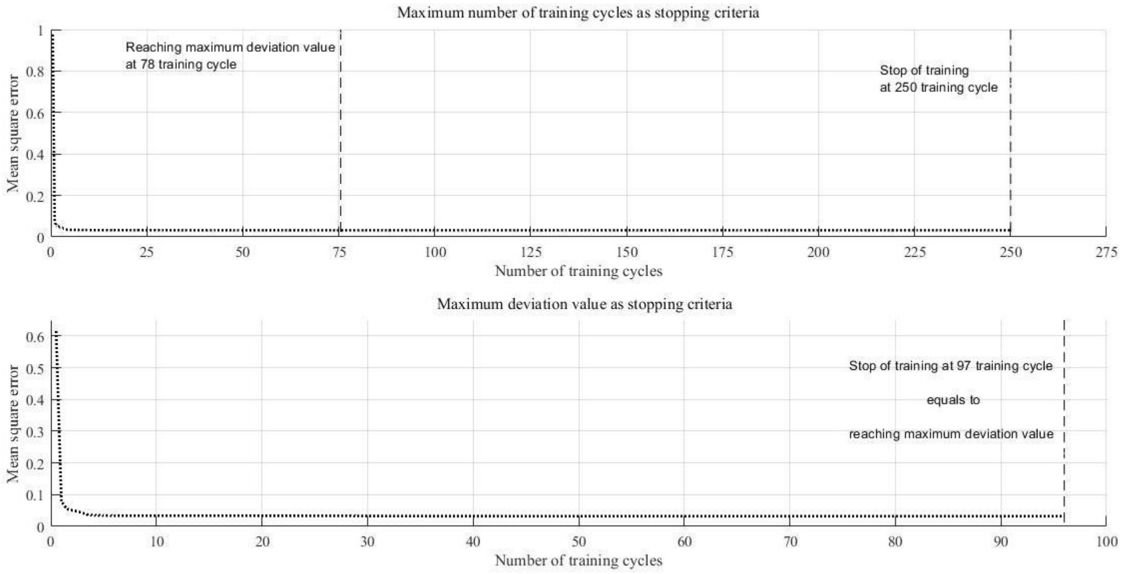

Fig. 3.5. Cases of training completion for the developed neural network model

After successful training of the network, the network is tested on the data, that were not in the training set. Then, to assess the model accuracy, the following accuracy parameters are calculated:

- maximum absolute error;

$$\Delta Y = \max[|Y_{out} - Y_{target}|] \tag{3.6}$$

Where Y_{out} are the values of the neural network output, Y_{target} are target parameter values.

- mean absolute error;

$$\Delta Y = avg[|Y_{out} - Y_{target}|] \tag{3.7}$$

- maximum relative error (in %);

$$\delta_Y = \max[\frac{|Y_{out} - Y_{target}|}{Y_{target}}] \tag{3.8}$$

- average relative error (in %);

$$\delta_Y = avg[\frac{|Y_{out} - Y_{target}|}{Y_{target}}] \tag{3.9}$$

- standard deviation;

$$MSE = \frac{\sum_{i=1}^{n}(Y_{out} - Y_{target})^2}{n} \tag{3.10}$$

- determination coefficient is a parameter that shows how the investigated model corresponds to the data. If the determination coefficient equals 1, the data is exactly described by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_{out} - Y_{target})^2}{\sum_{i=1}^{n}(Y_{out} - \frac{\sum_{i=1}^{n} Y_{out}}{n})^2} \tag{3.11}$$

According to the listed parameters, we conclude that our model can be used to solve the problem. If the model has low accuracy, it is trained again, or a different model architecture is used.

4. APPLICATION OF THE NEURAL NETWORK MODEL

According to the results obtained in this study, the neural network model can be used to calculate the component composition of natural gas, where the composition is measured by physical parameters. To check how accurate the method is, it was tested on a data sample. This sample included data that was not used for training the neural network model. The test sample includes 2218 gas mixtures, the ranges of its components are shown in Table 4.3. To calculate the physical chemical parameters of the simulated sample, we used NIST REFPROP software [8]. To calculate the speed of sound, the GERG-2008 gas state equation was used at standard temperature and pressure conditions. To calculate the thermal conductivity of the gas mixture, models for individual components and extended models for the corresponding states were used; the models are implemented in NIST REFPROP.

Table 4.3. Ranges of components molar fractions for test sample

Component	Molar fraction,

	%
Methane	80,5 – 99,5
Ethane	0 – 4,5
Propane	0 – 2
Carbon dioxide	0 – 9,5
Nitrogen	0 – 9,5

Table 4.4 and fig. 4.6 show how accurately the model predicts the natural gas composition of the test sample. Carbon dioxide deviation was set to zero, because the content of this component in the gas sample is known since the non-dispersive absorption of infrared radiation (NDIR) makes it easy to measure carbon dioxide molar fraction. Calculation of the parameters, as well as constructing, training, and testing the neural network model was implemented by means of the Matlab 2018a package with the Deep Learning Toolbox [6].

Table 4.4. Accuracy of the test sample composition prediction using neural network model

Component	Maximum absolute error, %	Mean absolute error, %	Standard deviation	Determination coefficient
Methane	0,77	0,33	0,35	0,9999
Ethane	0,38	0,16	0,16	0,9999
Propane	0,27	0,12	0,12	0,9999
Nitrogen	0,23	0,09	0,11	0,9999
Carbon dioxide	0	0	0	1

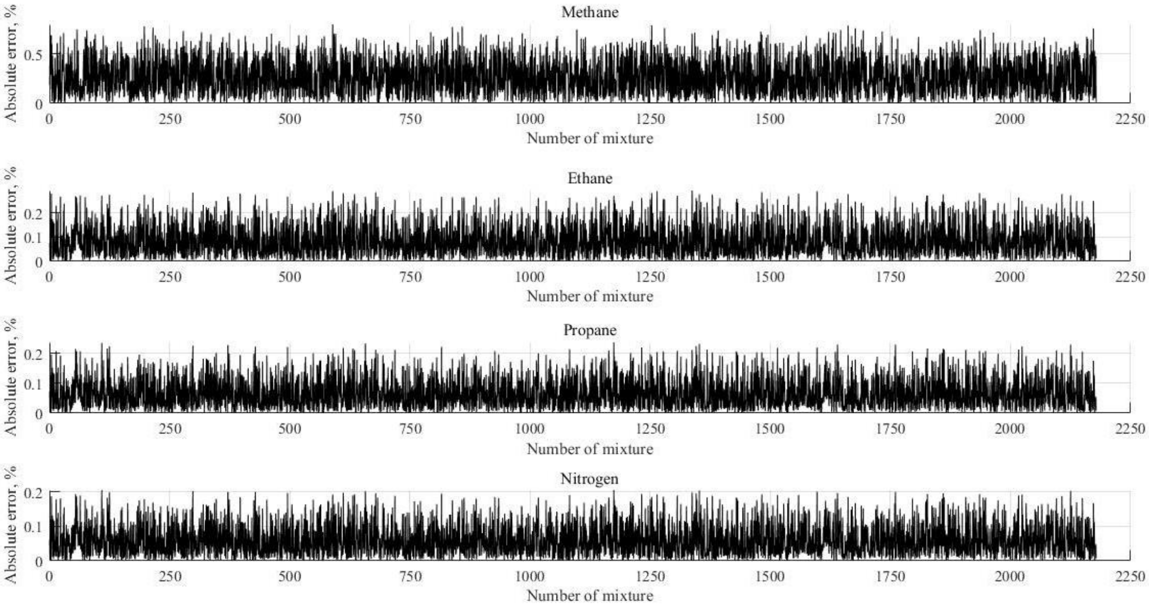

Fig. 4.6. Neural network prediction deviation (in absolute scale) for the test sample

5. CONCLUSION

The proposed conception of distributed data gathering system for natural gas composition analysis was used to determine gas composition by values of the gas measurable parameters. However, expansion of such systems is limited by huge complexity of distributed software debugging [10]. The main tasks of further research are to develop a neural network model with a more complicated architecture, to optimize the number and set of input measurable parameters to make this model more accurate.

ACKNOWLEDGEMENTS

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. Altfeld K., Schley P. (2012) Development of natural gas qualities in Europe, *Heat processing*, **3**, 77-83.
2. Dorr H., Koturbash T., Kutcherov V. (2019) 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.
3. *Dynament Infrared Gas Sensors Website* [Online]. Available: https://www.dynament.com.
4. Koturbash T., Bicz A., Bicz W. (2016) New instrument for measuring velocity of sound and quantitative characterization of binary gas mixtures composition, *Measurement Automation Monitoring*, **3**, 254-258.
5. Koturbash T.T., Brokarev I.A. (2018) Metod opredeleniya svoystv i sostava prirodnogo gaza po izmereniyam ego fizicheskikh parametrov [The method of determination of the properties and the composition of natural gas by measuring of its physical parameters]. *Sensors & Systems*, **6**(1), 43-50, [in Russian].
6. *Matlab 2018b Software* [Online]. Available: https://www.mathworks.com.
7. *Modcon Systems LTD Website* [Online]. Available: https://www.modcon-systems.com.
8. *REFPROP Software* [Online]. Available: https://www.nist.gov/srd/refprop.
9. Stepin U.P., Trahtengerts E.A. (2007) *Kompyuternaya podderzhka upravleniya neftegazovymi tekhnologicheskimi protsessami i proizvodstvami.Kniga 1. Metody i algoritmy formirovaniya upravlencheskikh resheniy* [Computer-aided management of oil and gas technological processes and industries. Book I. Methods and algorithms of managerial decisions]. 221-223, [in Russian].
10. Vaskovskii S.V. (2009) Sredstva otladki raspredelennykh mikroprotsessornykh kompleksov realnogo vremeni [Ways of debugging distributed microcomputerized real time systems]. *Sensors & Systems*, **1**, 44-45, [in Russian].
11. *Xensor Integration Website* [Online]. Available: http://www.xensor.nl