ISSN 2664-9969

UDC 004.8
DOI: 10.15587/2312-8372.2019.188743

Mishchuk O.

DEVELOPMENT OF THE METHOD OF FORECASTING THE ATMOSPHERIC AIR POLLUTION PARAMETERS BASED ON ERROR CORRECTION BY NEURAL-LIKE STRUCTURES OF THE MODEL OF SUCCESSIVE GEOMETRIC TRANSFORMATIONS

У роботі описано важливість удосконалення існуючих та дослідження нових алгоритмів прогнозування параметрів забруднення навколишнього середовища для поліпшення якості моніторингу навколишнього середовища. Оскільки організація та управління виробництвом вимагають розробки нових підходів до проблеми контролю та управління промисловими джерелами викидів шкідливих речовин на основі нових інформаційних технологій. Одним із найбільш проблемних місць систем контролю та управління якістю повітря на виробництві є розробка пришвидшених перспективних алгоритмів прогнозування забруднення повітря. Дані алгоритми повинні враховувати ситуативні зміни в розподілі даних та не вимагати перенавчання засобів прогнозування параметрів забруднення атмосферного повітря. З появою нейроподібних структур виникла потреба у їх дослідженні, в тому числі для задачі прогнозування параметрів забруднення атмосферного повітря. Об'єктом досліджень є нейроподібні структури моделі послідовних геометричних перетворень. Запропоновано метод прогнозування параметрів забруднення атмосферного повітря на основі корекції похибки за допомогою комітету нейронних структур різних типів. В ході дослідження проаналізовано три методи прогнозування параметрів забруднення атмосферного повітря: узагальнену регресійну нейронну мережу, радіально-базивну функцію та нейроподібну структуру моделі послідовних геометричних перетворень. Виконано комбінування згаданих методів та порівняно результати виконання трьох методів. Експериментально визначено, що прогнозування параметрів забруднення атмосферного повітря на основі корекції похибки за допомогою комітету нейроподібних структур моделі послідовних геометричних перетворень функції забезпечує зменшення похибки прогнозування. Зменшення на 7 % загальної регресійної нейронної мережі та на 2,6 % відносно радіально-базичної мережі з розширенням загальною регресійною нейронною мережею. Отримані результати забезпечують підвищення надійності та швидкості прогнозування параметрів атмосферного повітря для підвищення якості моніторингу викидів шкідливих домішків на виробництві та для прийняття управлінських рішень щодо природоохоронних дій.

Ключові слова: *атмосферне повітря, нейроподібна структура, головні компоненти, похибка прогнозування.*

Received date: 24.09.2019
Accepted date: 15.10.2019
Published date: 30.12.2019

1. Introduction

Environmental monitoring is an intelligent system with a wide variety of modules that provides the collection and processing of information obtained in the selected space-time field, further interpretation of the material, modeling, forecasting and management decisions [1].

The purposes of environmental monitoring are identification of potential hazards, development of measures to protect and prevention of the occurrence of critical situations, harmful or dangerous to human health and the existence of living organisms. The main tasks of environmental monitoring are monitoring the state of the environment, assessing and forecasting its status, determination of the degree of anthropogenic impact on the environment and identification of factors and sources of impact [2].

The objects of environmental monitoring are the environment and its elements, in particular, air. The environmental impact of the technogenic impact on the air environment is evaluated according to its monitoring data [3]. Monitoring of atmospheric air pollution gives a characteristic for a certain period of time the ecological state of the air environment and a forecast of the development of this state.

The function of predicting the state of the atmospheric air is recognizing the trends and logic of development of change of this state. Therefore, forecasting is intended to create a basis for making optimal management decisions. The information block provides the generation of environmental information needed to fully substantiate management decisions, for example for traffic restriction decisions on an area where atmospheric air pollution is predicted to increase [4].

Existing methods of forecasting of time sequences of pollution parameters, for environmental control with advancement, are not effective because of insufficient accuracy. Also, existing approaches to forecasting that are built on heuristic methods [5] and statistics algorithms [6] are not sufficiently accurate and therefore do not provide reliable recommendations for atmospheric air prediction. When using traditional neuro-paradigms, the entire time sequence is predicted, and it is not possible to eliminate the negative effects of random fluctuations and overall accuracy may be low. And situational changes in the data distribution require retraining of atmospheric pollution parameters.

The presence of a large number of man-made sources of danger, caused by the functioning of industrial production, containing many permanent sources of air pollution, poses a real threat to humans and the environment [7]. The quantitative environmental risk estimates obtained for a number of large industrial industries are quite high even for normal operation modes [8].

It is important to establish a maximum permissible concentration (MPC) for pollutants in order to use these standards when assessing damage and limiting impacts on natural objects. Air emissions from stationary sources are regulated for the most widespread and dangerous pollutants. The list of atmospheric air pollution parameters is established by the Cabinet of Ministers of Ukraine and is reviewed at least once every five years [9].

Analyzing and evaluating the state of the air environment is particularly important for choosing optimal management decisions, but they are based on the use of information that reflects present and past states. This is usually not enough to formulate a strategy, so trends in atmospheric air pollution must be taken into account to identify issues that may be encountered in the future. The determination of atmospheric pollution trends is a responsible and complex process, especially in conditions of environmental instability [10].

There are many approaches to forecast atmospheric air pollution using time series using artificial neural networks. These include multilayer perceptrons, neural networks of radial basis functions, recurrent neural networks, etc., all of which are based on the ability to approximate nonlinear functions [11].

For example, in [12] the authors investigate a hybrid statistical-logistic model of carbon monoxide prediction. In [13], the authors have developed predictive models of air pollution based on statistics using two neural network architectures: a multilayer perceptron and a nonlinear autoregressive exogenous network. The study [14] deals with the comparison of the seasonal autoregressive integrated moving average, artificial neural network and three models of fuzzy time series using the mean absolute error and the mean square error. It has been found that the accuracy of neural network prediction models is higher than that of other statistical models, but needs to be refined.

Therefore, there is a need to create more accurate neural network prediction algorithms which will take into account a large amount of environmental monitoring data and will require less time in application mode for use on mobile devices and controllers. With the development of computational intelligence and the emergence of neural structures, it is important to develop more advanced models, algorithms, and higher-speed forecasting tools of atmospheric pollution by high-speed neural structures. Thus, *the object of research* is neural-like structures of a model of successive geometric transformations.

The aim of research is development of a high-speed method for predicting the parameters of atmospheric air pollution based on neural-like structures of the model of sequential geometric transformations and to compare the proposed method with the statistical methods (based on General regression neural network).

2. Methods of research

Pollution prediction using different performance models can be divided into three types: potential forecasts, statistical models and numerical models [15]. For different elements, it is divided into pollution potential prediction and concentration prediction. The concentration forecast directly predicts the concentration of pollutants in a particular area.

Air pollution forecasting models can be divided into parametric and non-parametric ones. The parametric model is determination of the parameters of the equations in the known model, and its output is uncertain. For example, models based on a large amount of historical data, such as regression, principal component analysis, etc., are usually parametric models. Therefore, one of the simplest forecasting models used in practice is also the regression model (trend model). The dependent variable is the indicator under study, and the independent variable is the time or observation number of the indicator. A trend is a mathematical description of a time trend. Trend forecasting is to substitute the required numbers in the future instead of the observation number (or time) [16].

The purpose of trend analysis is to decompose the time series into principal components, measure the evolution of each component in the past, and extrapolate it into the future. The time mark by which future values of the time series are to be called is called the forecast horizon, which determines the size of the time interval expressed in the units of forecast (hours, days, months, etc.) for which the forecast is being constructed. Depending on the forecast horizon, the task of forecasting the time series is usually divided into the following categories of urgency: short-term forecast – from one day to one month; medium-term forecast: one month to one year and long-term forecast more than a year ahead [17].

The study considers a time series prediction algorithm based on a committee of the neural-like structures of various types including the Sequential Geometric Transformation Model (SGTM) [18]. Because of the aim to develop high-speed neuroimaging methods of improved accuracy in the prediction of atmospheric air pollution parameters. The SGTM neural-like structure in time series modeling provides an automatic decomposition of the time sequence into the following components: trend (trend of change) and oscillations of different frequencies, where the sum of all selected components is equal to the count of the time series [19]. Given the specificity of time series describing changes in atmospheric air pollution parameters over time, the periodic oscillations caused by many factors are virtually unpredictable. The volume of the input layer is chosen equal to the dimension of the input vector x. The number of output neurons is determined by the number of projected periods. The selection of the number of neurons in the hidden layer is determined practically

ISSN 2664-9969

by the method of selection with the onset of the most reliable result.

Increasing the accuracy of multi-step forecasting remains an important task as it increases the forecast horizon. The feature of the data is the nonlinearity of the response surface of the contamination parameters, as well as the limited number of samples that can be used for machine learning. So it is not possible to use nonlinear networks of high complexity, such as multilayer perceptron, radial basis function (RBF) networks with many centers [20].

Consider three possible options for improving accuracy, based on the principles of allocation of global and local components using the SGTM neural-like structure in auto social mode:

1. The training matrix contains vectors with inputs and output values:

$$X_{i,1},...,X_{i,n} \to Y_i.$$

2. The use matrix contains vectors with only input components:

$$X_{u,1},...,X_{u,j},...,X_{u,n}.$$

3. The test matrix contains the source components for the calculation of functional errors only and is not used in training.

To do preprocessing of data, perform scaling components of all vectors:

– for each column of data in the sample train find:

$$\max_{abs}(X_j), \quad \max_{abs}(Y_j);$$

– all components of the train, use, and test matrices change with formulas:

$$x_{i,j} = \frac{X_{i,j}}{\max_{abs}(X_j)}, \tag{1}$$

$$x_{i,j} = \frac{X_{u,j}}{\max_{abs}(X_j)}, \tag{2}$$

$$y_i = \frac{y_i}{\max_{abs}(Y_j)}, \tag{3}$$

where $i = \overline{1,N}$, $u = \overline{1,M}$.

2.1. Method 1 – Based on General Regression Neural Network (GRNN). For random k vector-point, the response (output) relative to the given (reference) points of the training sample has searched. It can be a point from the use or test matrix, and point from training matrix, for which the response is searched relative to other train points defined as anchor points.

The Euclidean distances from a given vector to all train reference vectors are calculated by the formula:

$$E_{k,i} = \sqrt{\sum_{j=1}^{n}(x_{k,j} - x_{i,j})^2}, \tag{4}$$

where k – number of matrices vector for which the forecasting is performed by GRNN; i – train vector reference number.

The next step contains calculating Gaussian functions from the Euclidean distances by the formula:

$$G_{k,i} = \exp\left(-\frac{E_{k,i}^2}{\sigma^2}\right), \tag{5}$$

where σ – Gaussian function span parameter, the selection of which is carried out by train vectors, excluding each of them from among the supporting ones.

The last step is to calculate the response by the formula:

$$y_k^{GRNN} = \frac{\sum_{i=1}^{N} G_{k,i} \cdot y_i}{\sum_{i=1}^{N} G_{k,i}}. \tag{6}$$

In some cases, the Gaussian function can be null, so protection against division by 0 is introduced:

$$\text{when} \ \sum_{i=1}^{N} G_{k,i} < 10^{-4}, \ \text{then} \ \sum_{i=1}^{N} G_{k,i} = 10^{-4}.$$

2.2. Method 2 – Based on Extended GRNN. For each input vector from training and use matrices additional component is introduced:

$$x_{i,1},...,x_{i,j},...,x_{i,n},y_i^{GRNN}.$$

For train vectors, each vector is considered, which is ejected from the N reference number, in turn, and it looks for the response y_i^{GRNN} relative to the N–1 reference ones.

The next step is to include it in the reference vectors and remove the next vector from the reference vectors and search for the next response.

2.3. Method 3 – Proposed Method. An error correction algorithm is implemented, which is based on the division of the error into errors of different characters, and consists of the following steps:

1. The outputs of the training sample are duplicated and the training of the neural-like structure is performed, allocating n main components, with the number of inputs n+1, resulting in the outputs being distorted.

2. Distorted outputs replace the initial outputs of the training sample after that training and applying is performed by the RBF network.

3. The same step is performed with the using linear SGTM neural-like structure.

4. It is provided that the prediction error by a nonlinear network is less than the prediction error by the linear SGTM neural-like structure, the prediction correction is performed by the following formula:

$$y_i = y_i^{RBF} - \alpha(y_i^{RBF} - y_i^{SGTM}), \tag{7}$$

where α – a proportionality factor that is experimentally selected.

3. Research results and discussion

The simulation methods are used to make local forecasts. When constructing forecast models, it is necessary to identify factors on which the forecast depends; find out their relationship with the predicted phenomenon; develop

an algorithm and programs for modeling environmental change under the influence of certain factors.

At the places of the greatest pollution of the atmospheric air, there are special measuring posts for monitoring the state of air pollution. There are 16 stationary posts in Kyiv (Ukraine), which monitor atmospheric air pollution with a sampling frequency of 6 days a week, 3–4 times a day. Borys Sreznevsky's Central Geophysical Observatory on the official site weekly shows partial results of environmental pollution monitoring [21]. A stationary observation post in Kyiv is selected for research, No. 7 – Bessarabia Square. The following parameters are measured here: dust, hydrogen chloride, nitrogen dioxide, sulfur dioxide, carbon monoxide, hydrogen fluoride, formaldehyde. The concentrations of impurities are shown in Table 1.

Table 1

Used data from the selected post

No.	Nitrogen	Fluoride	Chloride	Formaldehyde	Carbon
1	0.08	0.001	0.037	0.006	1.3
2	0.08	0.001	0.043	0.002	1
3	0.13	0.003	0.054	0.013	2.4
4	0.09	0.001	0.041	0.009	1.2
5	0.15	0.003	0.059	0.011	1.5
6	0.08	0.001	0.037	0.007	1.9
7	0.16	0.002	0.054	0.012	2
8	0.11	0.001	0.033	0.004	0.9
9	0.16	0.003	0.051	0.014	1.1
10	0.12	0.001	0.041	0.013	2.1
…	…	…	…	…	…
135	0.08	0.001	0.028	0.002	0.9
136	0.14	0.003	0.056	0.007	2.5
137	0.1	0.001	0.049	0.002	2

Before performing each of the methods, a trend is extracted from the data obtained. This process has consisted of the following steps:

1. Creating separate value matrices for each metric count, using sliding time windows, where the first n numbers of a row (x_1, x_2, x_3,..., x_n) are the input values of the neural-like structure and the last number (Y) is the desired output of a neural network whose value is duplicated from the last input value (x_n).

2. Extraction of one principal component and training of the linear SGTM neural-like structure in created samples with duplicates as additional inputs.

3. The outputs obtained during training are again divided by the sliding time window method and then divided into training and test samples.

Obtained training and testing matrices are used by the compared methods, the result of forecasting by which are presented in Table 2.

Table 2

Forecasting results by used methods

MAPE	Method 1	Method 2	Method 3
Testing error	2.82 %	2.69 %	2.62 %

Note: MAPE – mean absolute percentage errors

From Table 2, it can be seen that Method 2, which provides linearization of the overall response surface by adding GRNN input into training matrix-vector, shows better results than Method 1, which is used to provide a high level of generalization for a small sample. Also mean absolute percentage errors (MAPE) can be seen on the Fig. 1.

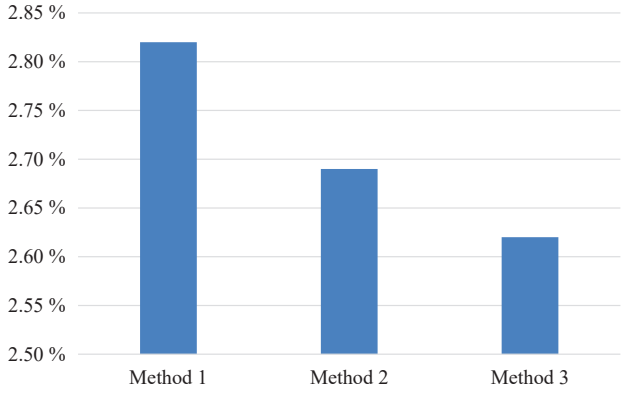

Fig. 1. Mean absolute percentage errors

Fig. 1 shows that the most accurate is proposed Method 3, which is based on errors in the formation of responses with errors of different characters, provided the overall accuracy of the combined structure is higher than for the linear application.

4. Conclusions

The novelty of the work is the use of the neural-like structure of the Successive Geometric Transformations Model to forecast the parameters of atmospheric air pollution with the allocation of the principal components for determining the trend and by performing the error correction based on the use of the neural-like structures of the Successive Geometric Transformations Model of different types.

During the research, it is experimentally proved that the proposed error correction method is more efficient than compared methods because it shows a forecast accuracy of 2.62 %, which are in 2.6 % and in 7 % better than other comparable methods for predicting air pollution parameters. Therefore, using the developed method, one of the environmental monitoring tasks can be more accurately performed than other comparable methods – prediction of atmospheric air pollution parameters for analysis and management decisions on environmental control.

The research results will be useful for short-term forecasting of atmospheric air pollution parameters and for widening the prediction horizon by reducing forecasting error.

References

1. Chaudhry, V. (2013). Arduair: Air Quality Monitoring. *International Journal of Environmental Engineering and Management, 4 (6),* 639–646.
2. Pope, C. A., Burnett, R. T., Thurston, G. D., Thun, M. J., Calle, E. E., Krewski, D., Godleski, J. J. (2004). Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution. *Circulation, 109 (1),* 71–77. doi: http://doi.org/10.1161/01.cir.0000108927.80044.7f

ISSN 2664-9969

3. Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S. W. et. al. (2013). The Changing Paradigm of Air Pollution Monitoring. *Environmental Science & Technology, 47 (20),* 11369–11377. doi: http://doi.org/10.1021/es4022602
4. Caubel, J. J., Cados, T. E., Preble, C. V., Kirchstetter, T. W. (2019). A Distributed Network of 100 Black Carbon Sensors for 100 Days of Air Quality Monitoring in West Oakland, California. *Environmental Science & Technology, 53 (13),* 7564–7573. doi: http://doi.org/10.1021/acs.est.9b00282
5. Xu, X., Ren, W. (2019). Prediction of Air Pollution Concentration Based on mRMR and Echo State Network. *Applied Sciences, 9 (9),* 1811. doi: http://doi.org/10.3390/app9091811
6. Rybarczyk, Y., Zalakeviciute, R. (2018). Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. *Applied Sciences, 8 (12),* 2570. doi: http://doi.org/10.3390/app8122570
7. Darbre, P. (2018). Overview of air pollution and endocrine disorders. *International Journal of General Medicine, 11,* 191–207. doi: http://doi.org/10.2147/ijgm.s102230
8. Gómez-Losada, Á., Pires, J. C. M., Pino-Mejías, R. (2016). Characterization of background air pollution exposure in urban environments using a metric based on Hidden Markov Models. *Atmospheric Environment, 127,* 255–261. doi: http://doi.org/10.1016/j.atmosenv.2015.12.046
9. Satish, U., Mendell, M. J., Shekhar, K., Hotchi, T., Sullivan, D., Streufert, S., Fisk, W. J. (2012). Is CO_2 an Indoor Pollutant? Direct Effects of Low-to-Moderate CO_2 Concentrations on Human Decision-Making Performance. *Environmental Health Perspectives, 120 (12),* 1671–1677. doi: http://doi.org/10.1289/ehp.1104789
10. Bai, Y., Li, Y., Wang, X., Xie, J., Li, C. (2016). Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. *Atmospheric Pollution Research, 7 (3),* 557–566. doi: http://doi.org/10.1016/j.apr.2016.01.004
11. Weizhong, Y. (2012). Toward Automatic Time-Series Forecasting Using Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems, 23 (7),* 1028–1039. doi: http://doi.org/10.1109/tnnls.2012.2198074
12. Gokhale, S., Khare, M. (2005). A hybrid model for predicting carbon monoxide from vehicular exhausts in urban environments. *Atmospheric Environment, 39 (22),* 4025–4040. doi: http://doi.org/10.1016/j.atmosenv.2005.04.010
13. Schornobay-Lui, E., Alexandrina, E. C., Aguiar, M. L., Hanisch, W. S., Corrêa, E. M., Corrêa, N. A. (2019). Prediction of short and medium term PM10 concentration using artificial neural networks. *Management of Environmental Quality: An International Journal, 30 (2),* 414–436. doi: http://doi.org/10.1108/meq-03-2018-0055
14. Rahman, N. H. A., Lee, M. H., Suhartono, Latif, M. T. (2014). Artificial neural networks and fuzzy time series forecasting: an application to air quality. *Quality & Quantity, 49 (6),* 2633–2647. doi: http://doi.org/10.1007/s11135-014-0132-6
15. Papanastasiou, D. K., Melas, D., Kioutsioukis, I. (2007). Development and Assessment of Neural Network and Multiple Regression Models in Order to Predict PM10 Levels in a Medium-sized Mediterranean City. *Water, Air, and Soil Pollution, 182 (1-4),* 325–334. doi: http://doi.org/10.1007/s11270-007-9341-0
16. Chowdhury, D. R., Sen, D. (2017). Artificial Neural Network Based Trend Analysis and Forecasting Model for Course Selection. *International journal of computer sciences and engineering, 5,* 20–26.
17. Caselli, M., Trizio, L., de Gennaro, G., Ielpo, P. (2008). A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model. *Water, Air, and Soil Pollution, 201 (1-4),* 365–377. doi: http://doi.org/10.1007/s11270-008-9950-2
18. Tkachenko, R., Izonin, I. (2018). Model and Principles for the Implementation of Neural-Like Structures Based on Geometric Data Transformations. *Advances in Computer Science for Engineering and Education*. Cham: Springer, 578–587. doi: http://doi.org/10.1007/978-3-319-91008-6_58
19. Izonin, I., Tkachenko, R., Kryvinska, N., Tkachenko, P., Greguš ml., M. (2019). Multiple Linear Regression Based on Coefficients Identification Using Non-iterative SGTM Neural-like Structure. *Lecture Notes in Computer Science*. Cham: Springer, 467–479. doi: http://doi.org/10.1007/978-3-030-20521-8_39
20. Mishchuk, O., Tkachenko, R., Izonin, I. (2019). Missing Data Imputation Through SGTM Neural-Like Structure for Environmental Monitoring Tasks. *Advances in Intelligent Systems and Computing,* 142–151. doi: http://doi.org/10.1007/978-3-030-16621-2_13
21. *Posterezhennia za zabrudnenniam atmosfernoho povitria v m. Kyievi*. Available at: http://cgo-sreznevskyi.kiev.ua/index.php?fn=lsza&f=lsza

Mishchuk Oleksandra, *Postgraduate Student, Department of Publishing Information Technologies, Lviv Polytechnic National University, Ukraine, e-mail: Oleksandra.myroniuk@gmail.com, ORCID: http://orcid.org/0000-0001-6823-985X*