International Journal of
Science and Engineering Investigations vol. 8, issue 91, August 2019
Received on August 19, 2019 ISSN: 2251-8843

The Accelerated Method of Filling Gaps in Data Using a Linear SGTM Neural-Like Structure

Oleksandra Mishchuk[1], Roman Tkachenko[2], Volodymyr Pohrebennyk[3]
[1,2]Department of Publishing Information Technologies, Lviv Polytechnic National University
[3]Department of Ecological Safety and Nature Protection Activity, Lviv Polytechnic National University
([1]oleksandra.mishchuk@gmail.com, [2]roman.tkachenko@gmail.com, [3]vpohreb@gmail.com)

***Abstract*–** In the article, the main methods of filling missing data have analyzed. The results of the comparison of missing data recovering methods in air pollution monitoring have presented. A method of filling the missing values in data, based on the use of artificial neural networks, is proposed. The choice of the neuro-like structure of the successive geometric transformations model for the problem of missing data reconstruction is substantiated. To reduce the use of computer resources, it has suggested using the neural-like structure of the Successive Geometric Transformations Model for training and the linear polynomial resulting from the training procedure for the application mode. The time of operation of the combined missing data recovering method in the environmental monitoring data of air pollution is experimentally determined. By comparison with the neural-like structure of the Successive Geometric Transformations Model for the problem of missing data reconstruction, it has proved that the combined method has a higher speed while maintaining high accuracy.

Keywords- *Neural Networks, SGTM, Impurity Concentrations, Filling the Omissions*

I. INTRODUCTION

Today, data science is an integral part of life that is constantly being used and developed in many different fields - business, social networks, medicine, video surveillance systems, engineering, geology, ecology, the stock market, etc. In recent years, there has been an increase in interest in using neural networks to solve a variety of applications. The use of neural networks has come into practice wherever the tasks of prediction, classification, recognition, or management are required [1].

This development is due to the following: neural networks - a powerful modeling method that allows reproducing complex dependencies. Also, a major factor in the development of this area is efficiency in use: neural networks have trained on examples, that is, on existing historical data, which had collected a lot [2].

During the learning phase, neural networks can detect complex relationships between input and output, and also generalize. The neural network user picks up the data and runs the training algorithm. Of course, the user must have a certain set of heuristic knowledge of how to select and prepare the data, select the desired network architecture, and interpret the results. However, the level of knowledge required for the successful use of neural networks is much lower than, for example, when using some methods of statistics [3].

II. OMISSIONS RECOVERING IN AIR POLLUTION AREA

The state of atmospheric air in Ukraine is unsatisfactory, and in some cities (Mariupol, Zaporizhzhia, Kryvyi Rih, Uzhgorod) it is quite threatening. The main sources of air pollution in Ukraine are industry - 65% and motor transport - 35%. The largest pollutants are coal-fired power plants. This represents about 27% of all emissions into the atmosphere. The remaining emissions are in ferrous and non-ferrous metallurgy. Fly ash, soot, and other impurities accumulate in the atmosphere. Flying dust contains silicon, calcium, magnesium, aluminum, iron, potassium, titanium, and sulfur. Vehicle exhaust contains oxidants, carbon monoxide, hydrocarbons, lead, and soot. About 60,000 Ukrainians die of dirty air every year. According to the State Statistics Service, in 1992 Ukraine was among the seven most polluted countries in the world, emitting 15.5 million tons of pollutants, and in 2017 they were reduced to 2.58 million tons [4].

In order to control the level of atmospheric pollution, in the places of accumulation of the largest emissions, measuring stations are installed. But not always measuring stations show complete observations. Gaps in the atmospheric air pollution monitoring data are caused by the failure of the measuring instruments or other reasons. As exceeding the emission limit values is dangerous for human life, there is a need to fill in the missing control parameters [5].

III. MATERIALS AND METHODS

Methods of filling missing data include a considerable number of algorithms: from simple approaches based on the substitution of certain values [6] to the use of modeling methods [7], which model the dependence of the estimates of the missing values on the observed ones. The methods of

filling omissions include: Nearest Neighbor method (replacing the missing value with the nearest information object); Zet algorithm (selecting each value to fill missing parameter not from the entire set of observations, but from some part of it) and Zetbraid (sequential selection of competent rows and columns and formation of a new matrix); *Barlett's method* (has two stages: changing missing data for the initial generated values in the first stage; conducting a covariance analysis of the target variable and constructing an indicator of completeness of observations for the target variable at the second stage); Resampling algorithm (an iterative method for changing rows with missing data by randomly selected rows from the full observation matrix); EM-estimation method (iterative procedure for optimization of some functional through analytic search for the extremum of function) [7].

The most promising for today are the methods of filling missing data based on data mining algorithms, which are able to identify internal patterns in the data and use them in the process of recovering missing values [8]. However, an analysis of the literary sources and recent trends in the field of filling missing data shows that it is extremely difficult to develop a universal model that is able to show the required results in different subject areas. Therefore, numerous papers offer models for specific subject areas, which include specialized algorithms for data processing based on machine learning methods and take into account the specifics of the information presented in them [7].

The choice of a neural network to solve the task of filling omissions in environmental monitoring data depends essentially on its architecture. Attention should be paid to performance metrics: network learning time, a number of parameters, and accuracy compared to other neural network models [9]. Performing experiments on selected data using different neural network models allow us to determine the optimal model for a particular task.

In previous researches [10,11], experiments were performed on the restoration of missed parameters of environmental monitoring of air pollution. The omissions occurred among the following parameters: non-methane hydrocarbons, titanium, benzene, tungsten monoxide and dioxide, indium oxide, temperature, nitrogen monoxide and dioxide, carbon monoxide. The selected parameters were taken from one of the stationary measuring points for air pollution in Italy [12]. Because the missing data were present for different parameters in the monitoring of air pollution, there was a need to train a large number of neural networks in different modes of application. The result of comparing gap filling with the following methods: regression algorithms: MLP, Random Forest, AdaBoost, SVC (with RBF kernel), SGD regressor, found that the neural-like structure of the Successive Geometric Transformations Model (SGTM) is optimal for the selected task [11]. It is the ICGP neural structure that has the most acceptable performance metrics, as it has the best training and application times and the highest precision.

However, there is always a need to further reducing the usage of time and memory (computer resources). Therefore, the study combines the neural-like structure of Successive Geometric Transformations Model for learning and linear polynomials to predict missing data. That is, instead of using a neural network to predict outputs, simple mathematical manipulations are used, which speed up prediction time by reducing memory costs.

Construction of the network (after selecting the input variables) comprises the following steps [13]:

1. Select the initial network configuration (for example, one intermediate layer with the number of elements in it).
2. Perform a series of researches with different configurations, while remembering the best network (in the sense of a control error). Several pieces of research should be performed for each configuration to ensure that the learning process has not reached the local minimum to obtain an incorrect result.
3. If a so-called “underdevelopment” is observed in the next experiment (the network does not produce an acceptable quality result), the usage of additional neurons in the intermediate layer or layers, or adding a new intermediate layer should be tried.

IV. PROPOSED OMISSIONS RECOVERY APPROACH

The algorithmic implementation of the filling missing data involves the following basic steps:

1. Teach the SGTM neural-like structure in the matrix of training data. The topology of the neural-like structure shown in Fig. 1 [13].

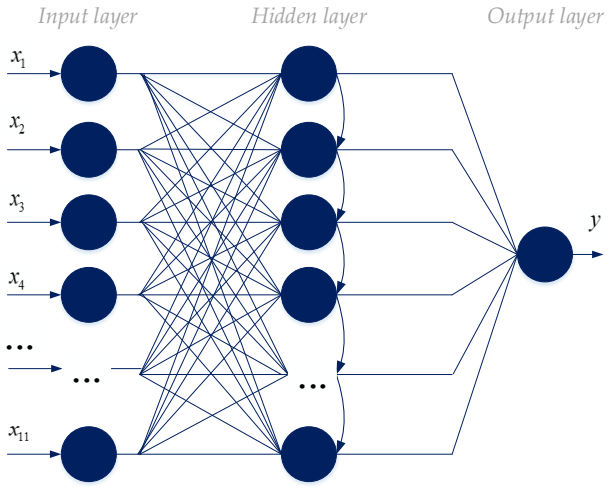

Figure 1. The topology of the SGTM neural-like structure

2. Apply the training network on the test matrix - diagonal matrix whose dimension is equal to the dimension of the input vector for training and one additional vector whose elements are zeros. It should be noted that the latter is the first vector of the matrix of test signals, which is shown in table 1. The purpose of this step is to obtain coefficients of a linear polynomial that will be used at the stage of application of the

method without using the neural-like structure of the Successive Geometric Transformations Model.

TABLE I. TESTING MATRIX TO PREDICT Y_N

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	...	X_i
0	0	0	0	0	0	0	0	0	...	0
1	0	0	0	0	0	0	0	0	...	0
0	1	0	0	0	0	0	0	0	...	0
0	0	1	0	0	0	0	0	0	...	0
0	0	0	1	0	0	0	0	0	...	0
0	0	0	0	1	0	0	0	0	...	0
0	0	0	0	0	1	0	0	0	...	0
0	0	0	0	0	0	1	0	0	...	0
0	0	0	0	0	0	0	1	0	...	0
0	0	0	0	0	0	0	0	1	...	0
...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	0	0	...	1

3. As a result of applying the neural-like structure of the Successive Geometric Transformations Model on the test matrix used in the previous step, the predicted outputs are obtained. These outputs are used as coefficients a_i for the linear polynomial (1) when solving the problem:

$$y_i = a_0 + a_1 * x_1 + ... + a_i * x_i \quad (1)$$

where coefficients $a_0 = y_0;\ \ a_i = y_i - y_0;$

x_i – input parameters of the test matrix for which executing the outputs are needed, that is, recovers the missing data.

Prediction of missed data by linear polynomials is done with software and written by the object-oriented programming language Java. Some pseudocode can be seen in Fig. 2.

```
long startTime = System.currentTimeMillis();

coefficients.add(inputCol.get(0));
for(int i = 1; i < inputCol.size(); i++)
{
        coefficients.add(inputCol.get(i) - inputCol.get(0));
}

final List<Double> predictedOutputs = new ArrayList<>();

final List<List<Double>> csvRows = new ArrayList<>();


for(int i = 0; i < prediction.size(); i++)
{
        Double predictedOutput = coefficients.get(0);
        final List<Double> row = prediction.get(i);
        for(int j = 0; j < row.size(); j++)
        {
                predictedOutput += coefficients.get(j + 1) * row.get(j);
        }
        predictedOutputs.add(predictedOutput);
}
```

Figure 2. Part of the program code to create a linear polynomial

Therefore, the recovery of the missing parameters in the environmental monitoring data of air pollution has performed by their prediction. Linear polynomials are used to reduce the prediction run time.

V. EXPERIMENTS AND RESULTS

In the research, the data from observations of atmospheric air pollution in Kyiv were used. Such observations are made by the Boris Sreznevskyj Central Geophysical Observatory. Systematic monitoring of the content of harmful substances in the atmospheric air is performed at 16 stationary posts with a sampling period of 6 days a week, 3-4 times a day.

During the observations, 20 contaminants are identified, but only nine are submitted for sharing on the official website [14]. Different posts show different amounts of measured impurities. Changes in published observations are made every week, so weekly measurements are monitored for the study and data are collected over two months. Information on various impurities and the state of atmospheric air pollution at the time of the research was given at four stationary posts: # 7 - Bessarabska Square, # 20 - Demiyivska Square, # 3 - Popudrenka Street, and # 5 - Nauka Avenue.

The study used data obtained at stationary post # 7 - Bessarabska Square. The concentrations of impurities are shown in Fig. 3.

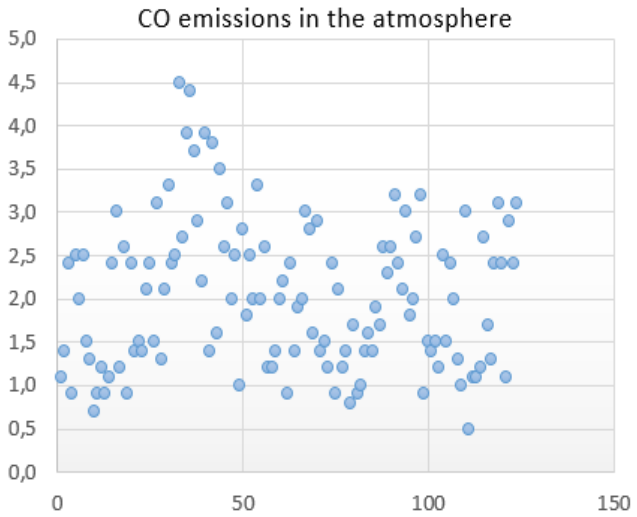

Figure 3. Impurity concentrations in these observations of Boris Sreznevsky Central Geophysical Observatory

At stationary post # 7, seven types of impurities are measured. The official site presents the following parameters of air pollution: Dust (Suspended solids); Sulfur dioxide; Nitrogen dioxide; Fluoride hydrogen; Hydrogen chloride; Formaldehyde; Carbon monoxide.

Detailed values of the impurities measured in the air are given in Table 2 [14].

TABLE II. PUBLISHED DATA FOR THE STATIONARY POST # 7

Suspended solids	Sulfur dioxide	Nitrogen dioxide	Fluoride hydrogen	Hydrogen chloride	Formal-dehyde	Carbon monoxide
0,180	0,060	0,060	0,002	0,047	0,003	0,7
0,180	0,065	0,050	0,001	0,053	0,003	0,9
0,180	0,084	0,150	0,003	0,070	0,006	1,2
0,180	0,078	0,080	0,002	0,059	0,003	0,9
0,180	0,081	0,250	0,003	0,085	0,004	1 ,1
0,180	0,081	0,080	0,002	0,049	0,004	2,4
0,220	0,085	0,110	0,003	0,079	0,007	3,0
0,180	0,084	0,070	0,001	0,052	0,003	1,2
0,190	0,090	0,130	0,003	0,063	0,007	2,6
0,180	0,080	0,080	0,002	0,056	0,006	0,9
0,180	0,077	0,040	0,002	0,050	0,009	1,4
…	…	…	…	…	…	…
0,190	0,083	0,050	0,002	0,056	0,005	2,1

There are maximum permissible concentrations (MPCs) for various atmospheric air pollutants. Since exceeding these limits seriously harms human health, air emissions are monitored. The values of maximum permissible concentrations (MPC) of pollutants in the atmospheric air of inhabited places are given in Table 3 [15].

TABLE III. MPC SUBSTANCES IN THE AIR

The name of the impurity	Maximum single MPC, mg / m3	Average daily MPC, mg / m3	Substance hazard class
Dust	0,5	0,15	3
Sulfur dioxide	0,5	0,05	3
Carbon monoxide	5,0	3,0	4
Nitrogen dioxide	0,20	0,04	3
Nitrogen monoxide	0,40	0,06	3
Fluoride hydrogen	0,02	0,005	2
Hydrogen chloride	0,2	0,2	2
Ammonia	0,2	0,04	4
Formaldehyde	0,035	0,003	2

As can be seen in Table 4, the hazard class of the substance is highest for carbon monoxide and ammonia. Ammonia in ambient air was not disclosed for the selected observation point. Carbon monoxide was only measured twice a day. Therefore, carbon monoxide was chosen for the task of recovering missing data by combining the neural-like structure of the Successive Geometric Transformations Model with linear polynomials.

From the previously reported data in Table 3, the training matrix was selected and the SGTM linear neural-like structure was trained. The training matrix included six inputs and one output. The inputs of the training matrix used all the pollution parameters except carbon monoxide. That only metric was used as output in the training matrix. After the study of the neural-like structure of the Successive Geometric Transformations Model, its application on the test matrix was performed to determine the coefficients of linear polynomials. The test matrix also consisted of six inputs and one output. The next step was to perform the prediction using linear polynomials and check the accuracy of the prediction. An excerpt of the algorithm for finding the mentioned errors is shown in Fig. 4.

```
FileWriter fw = new FileWriter(outputPath);
CSVWriter csvWriter = new CSVWriter(fw, SEPARATOR);

final List<Double> avg = new ArrayList<>();
for(int i = 0; i < 5; i++)
{
        avg.add(0.0d);
}

for (int i = 0; i < predictedOutputs.size(); i++)
{
        final Double a = error.get(i).get(0);
        final Double b = predictedOutputs.get(i);

        final List<Double> row = new ArrayList<>();
        row.add(a);
        row.add(b);

        final Double diff = b - a;
        row.add(Math.abs(diff));
        avg.set(2, avg.get(2) + Math.abs(diff));

        row.add(diff * diff);
        avg.set(3, avg.get(3) + diff * diff);

        row.add(Math.abs(diff / a));
        avg.set(4, avg.get(4) + Math.abs(diff / a));

        writeRow(csvWriter, row);
}

for(int i = 0; i < avg.size(); i++)
{
        avg.set(i, avg.get(i) / predictedOutputs.size());
}

avg.set(3, Math.sqrt(avg.get(3)));
avg.set(4, avg.get(4) * 100);

writeRow(csvWriter, avg);

csvWriter.flush();
csvWriter.close();
```

Figure 4. An excerpt of the algorithm for finding the mentioned errors

The accuracy of using the described approach to fill missing data by combining the neural-like structure of the Successive Geometric Transformations Model and the linear polynomials was verified by finding the errors shown in Fig. 5.

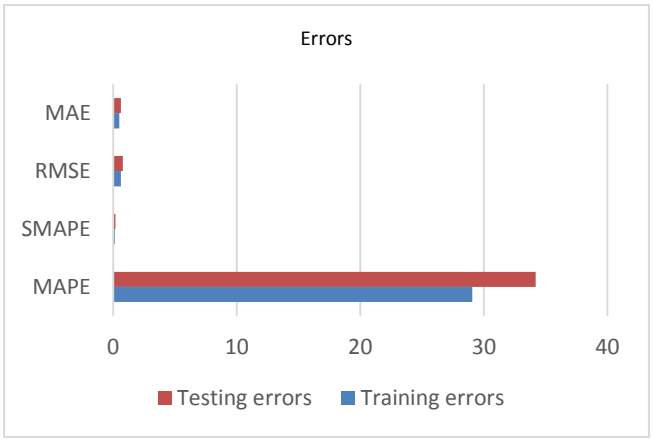

Figure 5. Training and application errors

Depicted as a graph, errors can be viewed in the form of Table 4.

TABLE IV. FOUND ERRORS

	Training errors	Testing errors
MAPE	29.06021857829584	34.19859081699813
SMAPE	0.12212937787374882	0.1692901356224873
RMSE	0.6167795105373846	0.7697291475748645
MAE	0.49801646310739806	0.6275186298323249

During experiments, it was determined that the prediction time using the non-iterative neural-like structure of the Successive Geometric Transformations Model was 0,007 milliseconds. Also, it was calculated that the time cost of predicting omissions by polynomials was 0,001 milliseconds. Time of predicting missing parameters are shown in Figure 6.

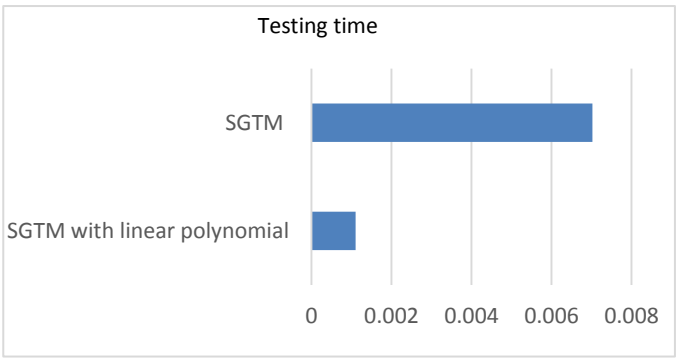

Figure 6. Training and application errors

Thus, the time of prediction execution by two methods is calculated and it is determined that filling missing data with the neural-like structure of the Successive Geometric Transformations Model combined with linear polynomials takes 7 times less time than the prediction using only the neural-like structure.

VI. CONCLUSIONS

In the research, the method of recovering missing parameters in environmental monitoring data of atmospheric air pollution with the help of their prediction is described. It is stated that the prediction of the missed parameters of air pollution control is made by combining a linear non-iterative neural-like structure of the Successive Geometric Transformations Model for learning and linear polynomials for prediction. Also, the study substantiates the use of the neural-like structure of the Successive Geometric Transformations Model to recover the missing data.

It is proven that this approach allows increasing the prediction speed while reducing the cost (saving) of the memory resource while maintaining the prediction accuracy.

REFERENCES

[1] Tkachenko, R., Izonin, I.: Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations. In: Hu, Z.B., Petoukhov, S., (eds) Advances in Computer Science for Engineering and Education. ICCSEEA2018. Advances in Intelligent Systems and Computing. Springer, Cham, vol.754, pp.578-587, 2019. https://doi.org/10.1007/978-3-319-91008-6_58

[2] M. Soley-Bori, "Dealing with missing data: key assumptions and methods for applied analysis", Technical Report, 2013, No. 4, pp. 1–20.

[3] Palamar M., Aleksander,M., Pohrebennyk V., Strembickyy M. Synthesis and optimization of neural network parameters for control of non-linear objects, Przeglad Elektrotechniczny Volume 90, Issue 5, 2014, pp 207-210. doi: 10.12915/pe.2014.05.47

[4] Karpinski M., Pohrebennyk V., Bernatska N., Ganczarczyk J., Shevchenko O. "Simulation of artificial neural networks for assessing the ecological state of surface water", 18th International Multidisciplinary Scientific Geoconference, SGEM 2018, Albena, Bulgaria, Volume 18, Issue 2.1, 2018, pp. 693-700. doi: 10.5593/sgem2018/2.1/S07.088

[5] V. Pohrebennyk, O. Korchenko, O. Mitryasova, N. Bernatska, M. Kordos, "An analytical decision support system in prognostication of surface water pollution indicators", 19th International Multidisciplinary Scientific Geoconference, SGEM 2019, Albena, Bulgaria, SGEM 2019, Vol. 19, Issue 2.1, pp 49-56.

[6] R. R. Andridge, J. A. Little, "A Review of Hot Deck Imputation for Survey Non-response", International Statistical Review, 2010, Vol. 78, Issue 1, pp. 40–64. doi: 10.1111/j.1751-5823.2010.00103.x

[7] E.-L. Silva-Ramirez, R. Pino-Mejías, M. López-Coello, M.-D. Cubiles-de-la-Vega, "Missing value imputation on missing completely at random data using multilayer perceptrons", Neural Networks, 2011, Vol. 24, Issue 1, pp. 121–129. doi: 10.1016/j.neunet.2010.09.008

[8] J. Honaker, G. King, M. Blackwell, "AMELIA II: a program for missing data", Journal of Statistical Software, 2011, Vol. 45, issue 7.

[9] I. Izonin, R. Tkachenko, N. Kryvinska, P. Tkachenko, M. Greguš, "Multiple Linear Regression based on Coefficients Identification using Non-Iterative SGTM Neural-Like Structure", In: Rojas I., Joya G., Catala A. (eds) Advances in Computational Intelligence, IWANN 2019, Lecture Notes in Computer Science, vol 11506, 2019, Springer, Cham, pp. 467-479.

[10] O. Mishchuk, R. Tkachenko, I. Izonin, "Missing Data Imputation Through SGTM Neural-Like Structure for Environmental Monitoring Tasks", In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in

Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing, vol 938, 2019 Springer, Cham, pp.142-151.

[11] I. Izonin, M. Greguš, R. Tkachenko, M. Logoida, O. Mishchuk, Y. Kynash, "SGD-based Wiener Polynomial Approximation for Missing Data Recovery in Air Pollution Monitoring Dataset", In: Rojas I., Joya G., Catala A. (eds) Advances in Computational Intelligence. IWANN 2019. Lecture Notes in Computer Science, vol 11506, 2019, Springer, Cham, pp 781-793.

[12] S. De Vito, M. Piga, L. Martinotto, G. Di Francia, "CO, NO, and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization", Sensors, and Actuators B: Chemical, 2009, Vol. 143, No. 1, pp. 182-191.

[13] Tkachenko, I. Izonin, P. Vitynskyi, N. Lotoshynska, and O. Pavlyuk, "Development of the Non-Iterative Supervised Learning Predictor Based on the Ito Decomposition and SGTM Neural-Like Structure for Managing Medical Insurance Costs," Data, vol. 3, no. 4, pp. 1-14, Oct. 2018.

[14] "Observation of air pollution in Kyiv". Accessed on: July 25, 2019. [Online]. Available: http://cgo-sreznevskyi.kiev.ua/index.php?fn=lsza&f=lsza

[15] "RD 52.04.186-89 "Guidelines for controlling atmospheric pollution". Accessed on: August 10, 2019. [Online]. Available: http://docs.cntd.ru/document/1200036406

How to Cite this Article:

Mishchuk, O. Tkachenko, R. & Pohrebennyk, V. (2019) The Accelerated Method of Filling Gaps in Data Using a Linear SGTM Neural-Like Structure. International Journal of Science and Engineering Investigations (IJSEI), 8(91), 154-159. http://www.ijsei.com/papers/ijsei-89119-20.pdf