Daily SO₂ Air Pollution Prediction with the use of Artificial Neural Network Models

Yasemin Gültepe

Department of Computer Engineering,
Kastamonu University, Kastamonu, Turkey

Ayşe Mine Duru

Department of Computer Engineering,
Kastamonu University, Kastamonu, Turkey

ABSTRACT

Although Artificial Neural Networks have been used for many years, its use in air pollution analysis has become widespread in 10 years. In this study, an Artificial Neural Network model was proposed to predication air pollution in Kastamonu province of Turkey. In this study as an example of Kastamonu province, Artificial Neural Network model was formed by using a pollution parameter (PM10) and 5 different meteorological factors (air temperature, air pressure, humidity, wind direction and wind speed) which were measured daily data during (2015-2018) period. It is aimed to propose refined model to predict the value of air pollution concentration (SO₂) after 24 hours by using this model. In other words, the air quality model is predicted. Artificial Neural Network is very successful compared to the new and classical statistical methods. Feed back-propagation algorithm has been used in all developed Artificial Neural Network model. The data set used in this study is divided into three subsets including training, validations and test data sets. The first 70% percent of the data set were used as the training subset, 15% of the data set used as the test set and 15% of the data set were used in validation set. The Mean Square Error (MSE) was measured for the performance of the network. It was observed that the developed ANN model was in agreement with the experimental results.

Keywords

Artificial neural networks, feed back-propagation algorithm, air pollution, Kastamonu.

1. INTRODUCTION

Air pollution is one of the biggest problems of nowadays. Air pollution, especially in the health of human beings, which adversely affect the health of the air and it is foreign matter in the air, the amount of density and reach above the normal. One of the most important steps in prevention of air pollution is the evaluation of the contamination event within a model [1]. Air quality models perform the transport and dispersion of pollutants in the atmosphere, chemical and physical transformations, removal processes, and simulations of pollutant concentrations and concentration areas at specific receptors, and prediction of concentrations over time [2]. Modeling and forecasting of air pollution, which can confirm very big problems in terms of health and negatively affect quality of life; it will contribute to timely measures and to the least extent of air pollution. Thus, it is possible to make appropriate decisions on controlling the behavior of different types of air pollutants based on model predictions and to prevent air pollution. Due to the continuous increase in air pollution in recent years, different technologies are used to predict air pollution. Among various computer techniques, the following are the main air pollution prediction model techniques [3]: Artificial Neural Networks (ANNs), Support Vector Machines, Fuzzy Logic, Hidden Markov Model,

Genetic Algorithm, Particle Swarm Intelligence, Hybrid Soft Computing Techniques, etc.

Since the first applications on air pollution models, ANNs have been the reliable method for achieving the targeted objectives and program goals [4]. ANN is a program designed to simulate the operation of a simple biological nervous system. They contain simulated nerve cells (neurons) and these neurons connect to each other in various ways to form a network. The learning algorithms commonly used in ANNs are also different from traditional computer algorithms. These algorithms contain the intuitive power of the human brain system. Thanks to their learning characteristics, ANNs can provide solutions to problems that are too complex for traditional techniques. Generally, the models created with ANN are used for statistical data analysis or system optimization [5]. In particular, ANNs have been applied to many problems of prediction and classification that depends on thousands or even millions of features [6].

In literature, many studies have been used on ANN to predict air pollution. In 2018, Alimissis et al. (2018), ANN model was evaluated [7]. Data from a real urban air quality-monitoring network located at Athens in Greece were used as the input layer and five regulated air pollutants (Nitrogen dioxide, Nitrogen monoxide, Ozone, Carbon monoxide and Sulphur dioxide) were used as the output layer. In most cases, ANN is found to be considerably superior to a decreasing spatial correlation between monitoring areas where network density is limited, especially in terms of air quality.

In 2005, Cigizoglu et al. (2005), November 1999-April 2003 period in Umraniye District on Istanbul's Asian Side, SO₂, NO and CO measurements were used. For air pollution parameter prediction, two ANN methods have been applied using Radial Basis Functions (RBF) and Feed Forward Error Back Propagation Method (FFBP). The results were compared with Multi Linear Regression (MLR). As a result, both methods provided significantly superior estimates compared to the traditional MLR method [8].

Although ANNs have been used for many years, its use in air pollution analysis has become widespread in ten years. In this study, it is aimed to prepare air pollution modeling program by using some meteorological parameters in prediction of air pollution with ANN models that give new and successful results in meteorology and environment applications by considering Kastamonu province of Turkey. In this study, the samples of Kastamonu province are measured as one air pollution particulate matter (PM10 (µg/m³)) and five different meteorological factors (air temperature, air pressure, humidity, wind direction and wind speed). It is aimed to prediction the value of the impurity parameter (Sulfur Dioxide (SO₂)) after 24 hours by using ANN models that are very successful in atmospheric applications and the results were compared with the actual values.

In this study, air pollution modeling in Kastamonu province is estimated from daily data received from Ministry of Environment and Urbanization Air Quality Monitoring Stations website. In the winter months, especially when the pollution is intensely experienced, taking precaution by estimating the air pollution due to meteorological conditions will contribute significantly to the reduction of the effect of air pollution. From this point of view, the importance of air pollution modeling and estimation studies is revealed.

2. ARTIFICIAL NEURAL NETWORKS

Inspired by human brain function, ANN is a system that can learn events by means of examples, and can generate and discover new information by using the information they have learned, so that the knowledge, experience and experience gained can be similar to those of people. They create their own experiences with the information they obtain from the samples and make decisions on similar issues. Thanks to these properties, classification, modeling and estimation applications are used in many areas [9].

The most important advantage of ANN according to conventional statistical techniques is that it can successfully model the functional structure in many different forms based on data in cases where the functional structure related to the data set cannot be determined precisely. In contrast to statistical methods, ANN does not require assumptions about distribution variables and assumptions. ANN is also capable of tolerating missing data of some variables. However, the dependence on the number of data during the ANN trainings, ie the increase in the performance of data and the difficulty in establishing the network model, can be seen as a disadvantage [1, 6].

One of the important areas where ANN is used is the estimation of future values. ANNs are powerful in terms of information processing and computing. It can be said that ANN has taken the power of computing and information processing from its parallel-scattered structure and its ability to learn and generalize. All these features show that the ANN is capable of solving complex problems. ANN is not linear. This has been the most important tool for ANN in solving particularly complex nonlinear problems [10].

Feed back-propagation algorithm was used in all of the ANN model developed in this study. Back-propagation algorithm is a supervised learning algorithm. In neural networks, dependent and independent variables are used as training data. Supervised training needs to hold back a set of data to be used to test the system after it has passed on its training. If data obtained from the independent variables in the test process enter the neural network and the values obtained are approximately the same as the dependent variables, the network is correctly trained [11].

3. MATERIAL AND METHOD

3.1 Data Set

The effects of air movements are closely related the geographic location and topographic structure of a settlement. Kastamonu city is located on the hill; the descent into the valley and the fact that it had to develop through the valley in time led to an increase in air pollution.

Today, with the developing industry, rapid urbanization and population growth, many cities of our country, like Kastamonu also has to be taken measures to prevent air pollution. Air pollution, which is seen as the problem of big

cities, has become the first priority problem in small cities today.

Air pollution measurements in Kastamonu province have been started regularly since 2002. The measurements of SO₂ and PM concentrations are examined. According to the seasonal changes results of the pollution measurement of the air were found that: a) Concentration of the pollution is increased in winter, but decreased at summer, b) Both concentration of the pollutants were higher than in the morning and in the evening, compared with other times in all days of the week, c) Concentrations are significantly decreased in week-end compared with other days of week. d) Concentration of the both pollutants were low during the 00:00-06:00, it was very high during the 06:00-10:00. After the 10:00, SO₂ concentration was continuously decreased, despite of the increase of dust [12].

The ANN model created in this study focused on the prediction of air quality of Kastamonu city center. Daily data during (2015-2018) period were taken from the station reports of Kastamonu province. KastamonuDataSet, the amount of components in the air; It was collected from the stalls from which it was obtained by measuring 24 times a day during (2015-2018) period.

Table 1. Statistical ratios for the parameters used in the ANN model

Parameter	Property			
	Unit	Mean	Std. Dev.	Median
PM10	µg/m ³	48.06	47.33	38.91
Air Temperature	°C	9.12	10.25	8.8
Air Pressure	mbar	967.31	200.68	1012.21
Humidity	%	68.98	23.09	76.23
Wind Direction	degree	195.02	121.12	237.1
Wind Speed	m/s	143	1.88	0.83

Statistical features of the data set are very important for data analysis and pre-processing. Statistical features are represented by the mean, standard deviation, variance, hypothesis tests etc. The statistical change interval of each meteorological parameter was examined. Table 1 is shown respectively statistical characteristics on the KastamonuDataSet as unit, mean, standard deviation and median values.

The six (6) parameters in Table 1 were used as the input parameters of the model to determine the next SO₂ concentration output characteristic. These data are sequentially presented as input data in the input layer of ANN model. In addition, the PM10 concentration of the day of meteorological data were used as input variables. Thus the possible effect of pollution parameters on the atmosphere was included in the model. The measurements of the stations may be missing at certain time intervals. The missing data is not included in the input data set. The entire data set for an average of 2500 days between October 4, 2015, and October 30, 2018, was used as a training set.

Before training is started, the collected data was divided into three subsets including training, validations and test data sets. This first subset is the training set, which is used to calculate the gradient and updating the network weights and biases. The second subset is the validation set, which is used to estimate prediction error for model selection. Finally, the test set is used to evaluate the performance of the model. The first 70% percent of the data set were used as the training subset to build the ANN model, 15% of the data set used as the test set and the remaining 15% of the data set were used in validation set of the developed ANN model.

Only one output, the daily average SO₂ parameter value was predicted. In other words, the effect of today's air pollution level on the air pollution level for tomorrow is included with this parameter. Air pollution after twenty-hours using the past five years parameters of data from KastamonuDataSet, in other words, air quality prediction model has been shown.

3.2 Methodology

In this study, ANN model was designed and developed in MATLAB 9.5 Neural Network Toolbox. In the entire developed ANN model, feed back-propagation algorithm on Levenberg-Marquardt algorithm has been used.

Fig 1: ANN model designed for the system

The general structure of ANN model is shown in Figure 1. ANN model presented here consists of three different layers of neurons: an input layer, a hidden layer and an output layer. There are six neurons in the input layer. The output layer also consists of one neuron representing the SO₂ concentration. Two hidden layers have been created on the ANN model. In the first hidden layer has been ten neurons. In the second hidden layer has been eight neurons. In hidden layer, learning methods and network parameters were changed to find the most suitable model. For the appropriate training of the developed ANN models hourly meteorological data are used.

As shown in Figure 1, the inputs of the network are information from other neurons or from the outside world. The aggregation function is a function that calculates the net input of the ANN model by collecting weights and input products. The weights indicate the importance of the information coming into an artificial cell and its effect on the cell. The information in the ANN model is kept in the weights of the connections in the network. In this study, the input from the neuron is multiplied with the weight coefficient and the b (threshold-bias) value is added to the ANN model, so that the net input is obtained. This aggregation function is expressed as Equation 1.

$$Net = \sum_{i=1}^n w_i x_i + b \quad (1)$$

In this study, sigmoid function was used as an activation function on network structure consisting of hidden layers. Experiments were performed to find the most appropriate number of neurons used in the interlayer by changing the learning methods and network parameters. The coefficient of determination (R^2) was used as the performance criteria. The Levenberg-Marquardt learning algorithm [13] provided the change in weights in feedback. It is the fastest and most powerful algorithm for training in medium-scale feed-forward

ANNs [14]. The data obtained from measuring 24 times a day during (2015-2018) period were used as input data with historical sequencing.

Sigmoid function is shown in Equation 2. This function is defined for y values exists between (0 to 1), and it corresponds to zero before x=0 axis and then converges to one. This function is particularly useful for the activation of neuron in an ANN.

$$f_{(Net)} = \frac{1}{1+e^{-Net}} \quad (2)$$

The data in the ANN must first be in a range and in a similar fashion with each other. If all data are in a range in value and not proportional to each other, that is, if the difference between the data is too large or too small, the resulting net inputs will be disproportionate. Therefore, it must be scaled at a specific interval to avoid computational errors.

Normalization is a scaling technique or a mapping technique or a preprocessing stage. There is a new range of data from an existing one range. Normalization can be very helpful for the prediction or forecasting purposes. Hence, it is necessary to maintain the vast diversity of estimates and to make the normalization technique closer to them. In this study, the minimum-maximum (Min-Max) normalization method is used. Min-Max normalization is a simple technique in which the technique can specifically fit data in a predefined boundary with a predefined boundary. X be a numerical variable that takes values $X=\{x_1, x_2, x_3, \dots, x_n\}$ in a data set records. Then the range of x in the data set is $[x_{min}x_{max}]$. In order for the data set to be scaled in the interval [0-1], the $X_{min}X_{max}$ range of the set is available and scaling can be made according to Equation 3.

$$X_{new} = \frac{X-X_{min}}{X_{max}-X_{min}} \quad (3)$$

Once the ANN has been structured and the training phase has been performed, the data taken for the test in the test phase is applied as input to the network. The performance of structured network is calculated by comparing the outputs obtained in this section with the actual outputs.

The training data were presented to ANN and the ANN model, which gives the smallest MSE value for the test data, was selected from the models where the network was learning. The Mean Absolute Percentage Error (MAPE) method refers to errors in percentages. This ensures that the errors are compared with each other. The lowest value as MAPE error value should give the recommended model with artificial neural networks. In this study, MSE was selected as performance function.

3.3 Results

Network performance is represented by the number of epoch (iteration). Network performance starts with a large value in the first epochs and is arranged to minimize this function by reducing the weights due to training. The performance chart consists of three lines because the input and target data are partitioned into three sets (train, valisation and test). In addition, a black dashed line is drawn, representing the network's best validation performance. Training stops when the green line representing the validation training set (network performance) intersects the black line. The MSE values obtained in each iteration are as shown in the figure below.

Fig 2: ANN Mean squared error performance

The figure 2, execution of the ANN model, was measured the MSE performance curve that occurs during the training of the network on the Levenberg-Marquardt learning algorithm. The training of the ANN was completed in 97. iteration. Speculation was ceased at the 103th iteration as there are no changes. The finest execution were acquired at 97th iteration, which the MSE value for this point was seen to be 0.0106 in ANN model in this study. As a result, instead of manual calculation of air pollution values, best validation performance value was estimated using the ANN model.

The figure 3 (a) to (d) consequences of regression graphs or training, testing, validation and normal of all. The graphs demonstrate a normal regression of $R=0.86918$ estimation, in other words, it is a regression value that implies that the predicted qualities are related to yield because the information in the graph is on the trend line. Similarly, it was independently approved for training, testing and validation.

Collation coefficients; For Training, Verification, Testing and All Data, respectively, it respectively was obtained as 0.87082 (a), 0.85697 (b), 0.87358 (c), and 0.86918 (d) for R. These results show that the model and the real data are in harmony with each other.

The ANN predictive model is used to verify network performance. The figure 3 shows the regression graphs showing the training results of our network. This regression line is taken as the correct base to draw the closest to the results. The closer this is to the right x (Target) = y (Output) line, the more successful it is. Multiplying the value displayed on each graph by 100 gives the success percentage of the process. The overall success rate of education, verification and testing in the chart in the lower right corner is between 85-87%. This high rate is the results of successful ANN model in this study.

Fig 3: Regression graphs of training, testing, validation and complete of all

4. CONCLUSION

The artificial neural network approach provides the easiness of obtaining the results of different systems by making changes in the desired parameters without the need to make an analysis by establishing an experimental system. It is also frequently used in the solution of systems with complex analytical solutions.

The main objective of this study is to estimate air pollution in Kastamonu province with the lowest error using ANN. ANN model was formed by using a PM10 and five different meteorological factors (air temperature, air pressure, humidity, wind direction and wind speed) which were obtained from measuring 24 times a day during (2015-2018) period. From the results, it can be concluded that ANN is 87% efficient with an error rate of 0.0106 at estimating air pollution in Kastamonu province.

Thus, the method presented in this study is a good estimate for air pollution in the city of Kastamonu province. In Matlab, this software can be used to estimate air pollution in daily life and to predict different meteorological events.

5. REFERENCES

- [1] Kunt, F. 2007. Hava Kirliliğinin Yapay Sinir Ağları Yöntemiyle Modellenmesi ve Tahmini, Selçuk University Graduate School of Natural and Applied Sciences, M.Sc. Thesis, Environmental Engineering Department, Konya, 2007.
- [2] Demirarslan, O. Çetin, Ş., and Ayberk, S. 2008. Hava Kirliliği Belirlemelerinde Modelleme Yaklaşımı ve Modelleme Aşamasında Karşılaşılabilecek Sorunlar, Environmental Problems Symposium, Kocaeli 2008.
- [3] Niharika, Venkatadri, M., and Rao, P.S. 2014. A survey on Air Quality forecasting Techniques, International Journal of Computer Science and Information Technologies, 5(1), 103-107, 2014.
- [4] Bonzar, M., Lesjak, M., and Mlakar, P. 1991. A neural network based method for short-term predictions of ambient SO₂ concentration in highly polluted industrial

- areas of complex terrain, *Atmospheric Environment* 27B (2), 221-230.
- [5] Yüksek, A.G., Bircan, H., Zontul, M., and Kaynar, O. 2007. Sivas İlinde Yapay sinir Ağları İle Hava Kalitesi Modelinin Oluşturulması Üzerine Bir Uygulama, *CU Journal of Economics and Administrative Sciences*, 8(1), 97-112.
- [6] Warner, B., and Misra, M. 1996. Understanding Neural Networks as Statistical Tools, *the American Statistician*, 50, 284-293.
- [7] Alimissis, A., Philippopoulos, K., Tzanis, C.G., and Deligiorgi, D. 2018. Spatial estimation of urban air pollution with the use of artificial neural network models, *Atmospheric Environment*, 191, 205-213, 2018.
- [8] Cigizoglu, H.K., Alp, K., and Kömürcü, M. 2005. Estimation of Air Pollution Parameters Using Artificial Neural Networks Advances in Air Pollution Modeling for Environmental Security. *NATO Science Series (Series IV: Earth and Environmental Series)*, 54, 63-75, 2005.
- [9] Öztemel, E. 2003. Yapay Sinir Ağları, Papatya Publishing, İstanbul.
- [10] Zhang, G., Patuwo, B.E., and Hu, M.Y. 1998. Forecasting With Artificial Neural Networks: The State of the Art, *International Journal of Forecasting*, 14, 35-62.
- [11] Patterson, D. 1996. Artificial Neural Networks, prentice Hall. New York.
- [12] Gemici, Z., Hoşafcıoğlu, D., and Tok, Ç. 2017. Kastamonu İlindeki Hava Kalitesi Sonuçlarının Değerlendirilmesi, VII. National Air Pollution and Control Symposium, 863-872, 2017.
- [13] Levenberg, K. 1944. A Method for the Solution of Certain Non-linear Problems in Least Squares. *Quarterly of Applied Mathematics*, 2(2), 164-168.
- [14] Ibrahimy, M.I., Ahsan, R., and Khalifa, O.O. 2013. Design and Optimization of Levenberg-Marquardt based Neural Network Classifier for EMG Signals to Identify Hand Motions, *Journal of the Institute of Measurement Science*, 13(3), 142-151.