Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Application of artificial neural networks model to predict the levels of sulfur dioxides in the air of Zamość, Poland

To cite this article: J Kujawska *et al* 2022 *J. Phys.: Conf. Ser.* **2412** 012005

Application of artificial neural networks model to predict the levels of sulfur dioxides in the air of Zamość, Poland

J Kujawska[1], M Kulisz[2] and Z Aubakirova[3]

[1]Faculty of Environmental Engineering, Lublin University of Technology, Nadbystrzycka 40B, Lublin, Poland
[2] Faculty of Management, Lublin University of Technology, Nadbystrzycka 38, Lublin, Poland
[3]School of Architecture and Construction, D. Serikbayev East Kazakhstan Technical University, 69 Protozanov Street, Ust-Kamenogorsk, Kazakhstan

j.kujawska@pollub.pl

Abstract. Air quality control and its prediction are particularly important for human health and life. Sulfur dioxide constitutes one of the air pollutants that play an important role in air quality pollution. An artificial neural network model was employed to forecast the levels of sulfur dioxide in the air of Zamość (Poland). The measured data of the meteorological station of Zamość in 2017-2019 were used for the model. Temperature (T), relative humidity (RH), wind speed (WS), wind direction (WD), SO_2, PM10, NO_2, NOx, CO, O_3, C_6H_6 were used as input parameters for building the neural network model. Regression value (R) and Mean Squared Error (MSE) were used to estimation the model. The results show that neural network is capable of predicting the sulfur dioxide levels in the air.

1. Introduction

Human life depends on air quality. The World Health Organization (WHO) reports that 33 cities in Poland had poor air quality [1]. Therefore, it is necessary to monitor and control air quality. The Environmental Protection Agency (EPA) lists six compounds that cause air pollution: SO_2, PM, O_3, C_6H_6, CO_2, NO_x. They are used for air quality assessment criteria [2].

Sulfur dioxide (SO_2) constitutes one of the most common air pollutants. It is a highly toxic gas, non-flammable with a strong suffocating odor. SO_2 is emitted into the atmosphere by industry, via combustion of sulfur-laden mine fuels, as well as from vehicle engine exhausts and domestic coal stoves. In the atmosphere sulfur dioxide oxidizes to sulfur trioxide. Next sulfur dioxide reacts with water to sulfuric acid (VI). These compounds cause acidification of soils, decrease their fertility, and inhibit plant growth. Long-term inhalation of sulfur oxides in humans causes damage to the respiratory tract and increased susceptibility to infection [3].

In Poland, the highest concentrations of sulfur dioxide occur in December and January, whereas the average value in autumn and winter is about 3.03 $\mu g/m^3$. The concentrations of sulfur dioxide in atmospheric air during the heating season are about 54% higher, compared to the summer season [4]. These values are lower than the permissible concentrations of sulfur dioxide specified in Directive 2008/50/EC of the European Parliament and of the Council (of May 21, 2008 on air quality and cleaner air for Europe). The aforementioned Directive specifies 350 $\mu g/m^3$ over a 1-hour and 125 $\mu g/m^3$ over a 24-hour as the permissible concentration of sulfur dioxide in the air for the protectian of human health and plants, while the alert level for the one-hour average concentration is 500 $\mu g/m^3$. Lawmakers also stipulated a "permissible frequency of exceeding the permissible level" of sulfur dioxide during the calendar year: it amounts to 24 times for hourly averaged results and 3 times for 24-hour results [5].

The variability of meteorological parameters, industrial development, complex chemical reaction and control measurments made it challenging to control outdoor air pollution. In Poland, the primary tools used in air quality control are the measurements conducted as part of state environmental monitoring networks. The second tool includes the models of pollutant dispersion and transformation in the air, as well as the models for predicting pollution levels [6].

Creating mathematical models for air quality forecasting is still a difficult task. Issues such as Big Data, Missing Data, Unreliable Data etc. are often encountered. Therefore, new models are still being developed using modern methods, e.g. machine learning.

Researches used various machine learning algorithms to forecast air quality. LASSO regression [7-9], Support Vector Machines (SVM) [10-13], Random Forest [14-17] and K-Nearest Neighbors (kNN)[18] models produce acceptable results to predict parameters of air quality. The analysis of the literature confirms that the mathematical models for air pollution forecasting show good quality of prognostic [19-21].

In air quality modeling studies, the trend of neural networks application can be observed in recent years. Researchers are proving that ANNs are an effective method for predicting air pollution, with very good precision in modeling in environmental engineering [2]. Thus, for example, Pérez and Trier (2001) modeled and predicted the nitrogen oxide levels using ANNs in the Santiago, Chile [22]. Ordières et al. (2005) created ANNs models to predict the level of PM2.5 [23].

This study investigated the predictive capabilities of artificial neural networks (ANNs) for predicting the airborne sulfur dioxide concentrations in Zamość (Poland). Most of the studies conducted so far have used meteorological data as input data for model building. In this study, in addition to the meteorological data and chemical air pollution data were used as input data. All data sets were taken from the air quality monitoring station located in Zamość at ul. Hrubieszowska, and 23300 hourly readings were used in the ANN models. The data for modeling came from the Reports of the Chief Inspectorate of Environmental Protection [24]. The regression value R and Mean Squared Error MSE were taken into account to evaluate model using the artificial neural network method.

2. Materials and methods

To predict the levels of sulfur dioxides (IV) in the air, the data from the air quality monitoring station of Zamość, Lublin Province in Poland was used. The meteorological data and air chemical pollution data read at every hour of every day in 2017-2019. The effects of the hour of measurement (h) and the month (M) were also taken into account, as they can have a strong influence on the level of sulfur dioxide in the air. The data on hours, months and wind direction are circular, therefore the changed two-dimensional sines and cosines have been added. The data used for the model are shown in Table 1.

Table 1. The data used for the model.

	Minimum	Maximum	Mean
SO_2 [$\mu g/m^3$]	0	45.8	3.71
PM10	1.3	358	20.73
NO_2	0	67.8	12.17
NO_x	1	142.3	15.7
CO	1	4.2	0.25
O_3	1	154.1	39.98
T	-17.6	35.7	11.3
RH	19	103	69.9
WS	0	19.8	6.01
WD	0.56	357	187.1

The first stage of the analyses conducted was choosing of optimal input parameters for creating the ANN model. For this purpose, correlation analysis was applied using the Statistica13 software. Next, modeling of the SO_2 concentration in the air was carried out using Matlab - Neural Network App software. The data used for modeling and the network diagram are shown in Figure 1

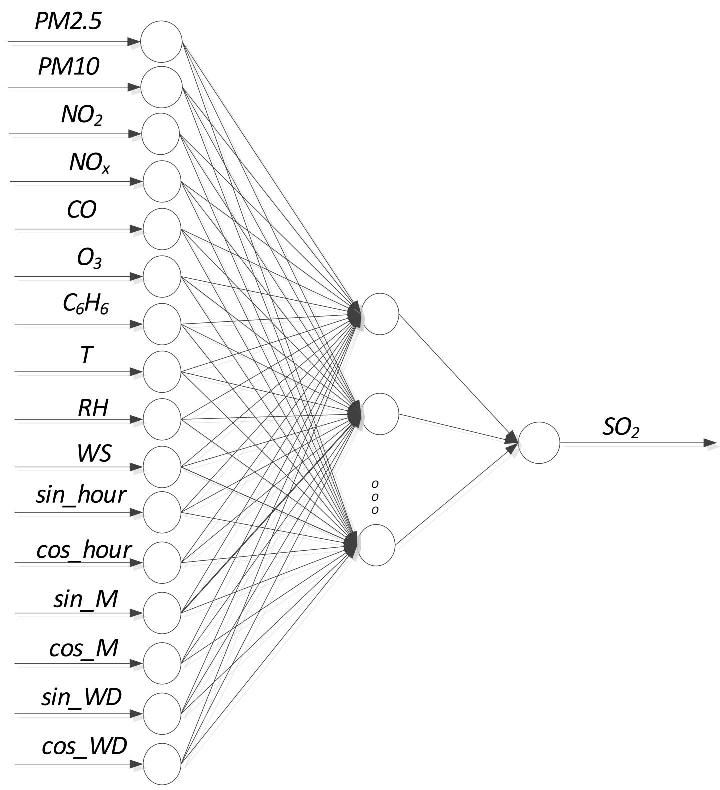

Figure 1. Scheme of ANN.

For modeling used a shallow network with 1 hidden layer, amount of neurons was selected experimentally in the range of 10 to 500, and three types of learning algorithms were tested for learning: the Levenberg-Marquardt algorithm (LM), Bayesian regularization algorithm (BR) and Scaled conjugate gradient algorithm (SCG). The dataset was divided in the ratio of 70%:15%:15% (learning set: test set : validation set). The quality of the network was determined by the regression value R, according to formula (1):

$$R(y', y^*) = \frac{cov(y', y^*)}{\sigma_{y'}\sigma_{y^*}} \qquad R\epsilon < 0; 1 > \qquad (1)$$

where σy′—the standard deviation of reference values, σy*—standard deviation of forecasting values.

In addition, the value of the Mean Squared Error (MSE), calculated according to formula (2), was taken into account:

$$MSE = \frac{1}{n}\sum_{n=1}^{n}(\hat{y}_i - y_i)^2 \qquad (2)$$

where y_i is the present value of the SO_2 level and $\hat{y}_i$ is the value of the SO_2 level for the i-th observation got from the model.

3. Results and discussion

In 2017-2019, the conditions for both the SO_2 permissible level and the frequency of exceeding this level were met in the Zamość (Table 1). There are few point sources of SO_2 emissions in the Zamość, as it is a minimally industrialized city. This is mainly due to the lack of major emitters and the eminently agricultural nature of the region. Only the ATEX Company in Zamość is subject to mandatory monitoring, whose SO_2 emissions in 2019 amounted to 176.8 tons. The primary source of sulfur dioxide emissions in the air was the energetic cambustion of sulfur-contaminated fuels, which has a direct impact on the seasonal variation in the concentrations of this pollutant throughout the year (Figure 2). The highest concentrations occur during the winter months. Average SO_2 concentrations in autumn and winter were about 3 times higher with respect to the average value of spring and summer.

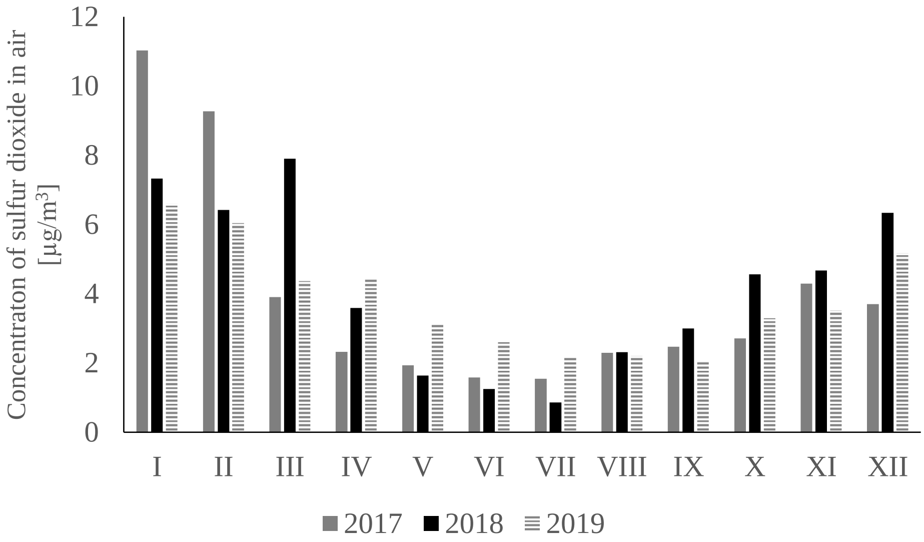

Figure 2. The average concentration of sulfur dioxide in each month according to Zamość station in the years of research.

Selection of optimal input factors affecting the SO_2 levels in the air constitutes a challenge in developing neural network models. For this purpose, a correlation matrix (Table 2) was created between the data from the monitoring station of Zamość and the level of SO_2 in the air. In interpreting the correlation coefficients, the criteria for the strength of the relationship between the data were adopted according to the classification proposed by Bam et al. (2011) [25]. The obtained correlation coefficients between the variables ranged from -0.046148 to 0.456916, indicating a lack of correlation between the data. Therefore, all analyzed input parameters were used to create an artificial neural network model (16): PM2.5, PM10, NO_2, NO_x, CO, O_3, C_6H_6, RH, T, WS, and sines and cosines circular data.

During modeling, networks with non-identical quantity of neurons in the hidden layer and using LM, BR and SCG learning algorithms were tested. The best network had 250 neurons, which was got in 33 iterations for the Levenberg-Marquardt algorithm. Figure 3 shown the structure of the network. The results of the learning process of the network divided into learning, testing and validation subsets and the values of MSE and R is shown on Table 3.

Figure 4 shows the gradient for a given iteratian of the validation set as a function of the number of consecutive increases in MSE, Mu (momentum) and the increase in validation error, where with six consecutive increases, model learning stopped, which occurred at the 33rd iteration. The top validation performance (3.646) was get at epoch 27 (Figure 5).

Table 2. Correlation matrix coefficients of SO_2.

Data from the monitoring station	Coefficient of correlation with the SO_2 level
PM 2.5	0.456916
PM10	0.451612
NO2	0.314275
NO_X	0.243569
CO	0.385297
O_3	-0.123145
C_6H_6	0.455515
T	-0.310550
RH	0.143635
WS	0.161119
SIN_HOUR	-0.054964
COS_HOUR	-0.046148
SIN_MONTH	0.278412
COS_MONTH	0.325760
SIN_WIND_DIR	0.031867
COS_WIND_DIR	-0.305331

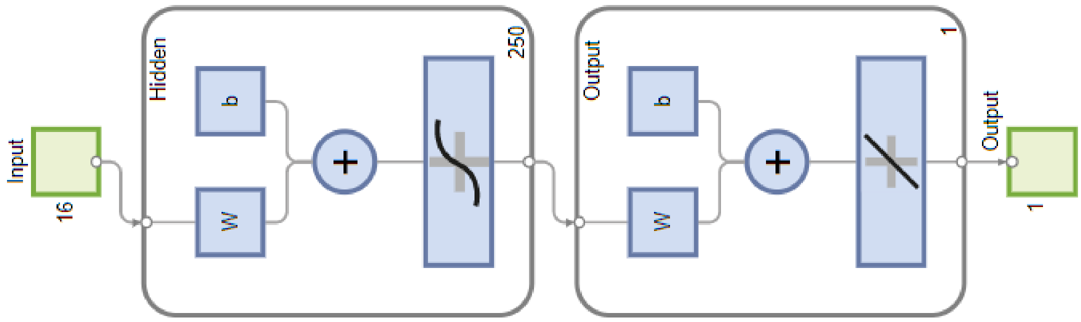

Figure 3. ANN structure.

Table 3. ANN regression statistic and MSE.

Data	Observations	Mean Square Error (MSE)	Regression (R) value
Training	16582	1.7328	0.9250
Validation	3553	3.6460	0.8460
Testing	3553	4.6587	0.8135

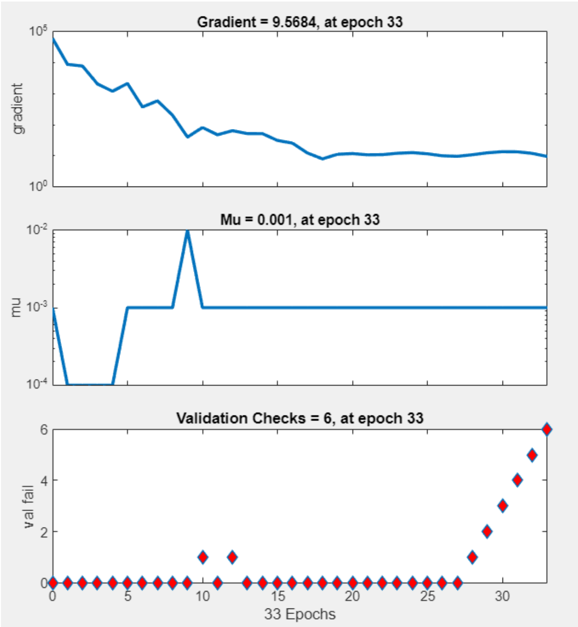

Figure 4. Early strategy for ANN model.

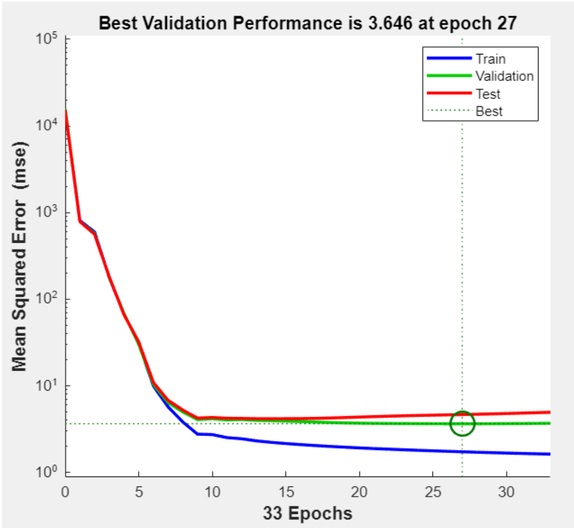

Figure 5. ANN teaching performance.

Figure 6 shows regression graphs or training, testing, validation and normal of all. The following regression statistics were obtained for each subset – learning R=0.925, validatian R=0.84603 and testing R=0.81346, and overall regression R=0.89382. These results exceed R>0.8, indicating a good network fit and a high level of agreement between measurement points. A comparison of the actual SO_2 data and predict SO_2 date is shown in Figure 7.

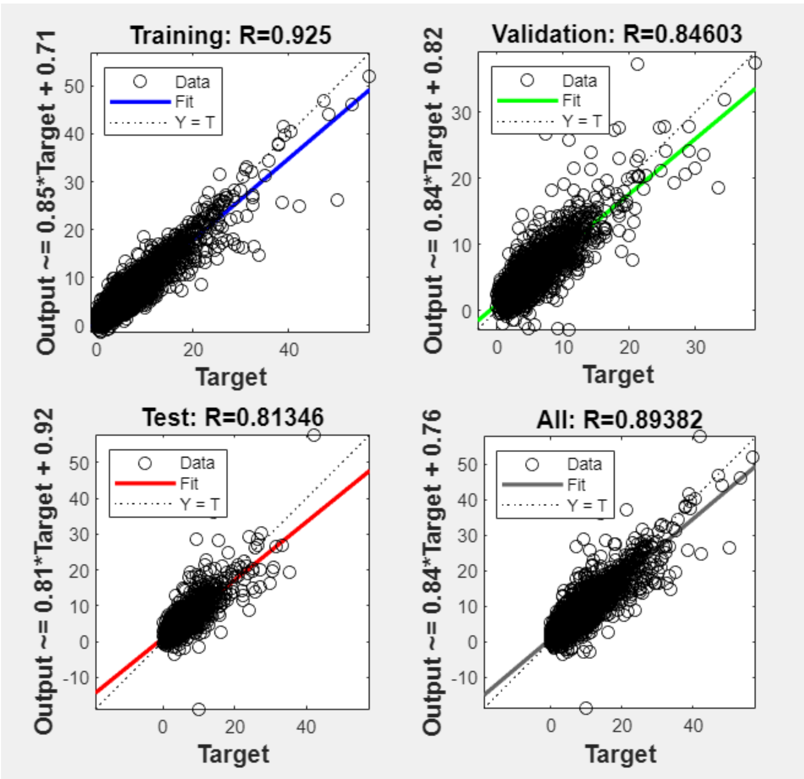

Figure 6. Regression graphs for individual and total set.

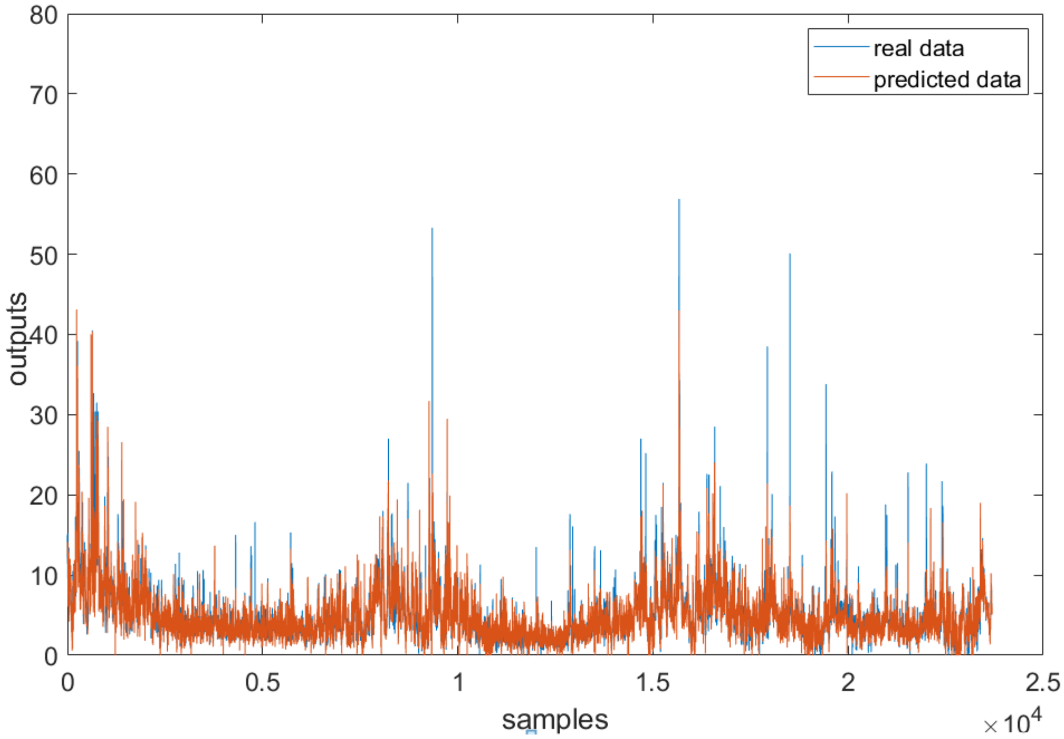

Figure 7. Comparison of the actual data and the predict data**.**

Taking into account the regression value R and the Mean Squered Error MSE, it can be concluded that the obtained network has an acceptable error rate and can be used as a reliable SO_2 predictor, which may be employed as a decision support tool in controlling and assessing outdoor air quality.

The results achieved in our study are not markedly from results another researchers. Researchers obtain artificial neural network models for predicting SO_2 with different qualities. Gültepe and Duru predicted the value of sulfur dioxide concentration in air (SO_2) after 24 hours by using artificial neural network model. They used metrological data of air: T, p, H, WD, WS from the measuring station of Kastamonu province in the period 2015-2018 as input. Their model had an overall regression equal to R=0.86918 [26].

Unnikrishnan and Madhu used ANN time series models to predict the levels of sulfur dioxides in the air. SO_2, NOx, NH_3, CO, particulate matter, T, RF (rainfall), RH, WD, WV (wind velocity) were used as inputs. Their model showed good quality evaluation characteristics with an MSE of 0.0115 and an R of 0.8979 [27].

In turn, Schams et al. predicted the SO_2 concentrations in Tehran's air using ANN models. Various parameters were selected to predict the daily SO_2 concentration, meteorological date, urban traffic data, urban green space data and time parameters. They obtained an MSE of 3.01 [28].

Chelani et al. predicted the sulfur dioxide concentration in three zones (industrial, commercial and residential) in Delphi and compared with the measured data. They obtained MSE for the prediction data equal to 0.76 for industrial sites, 0.72 for commercial sites and 0.67 for residential sites in Delphi [29].

This study demonstrates the importance of creating multiple ANN models to predict the sulfur dioxide levels in the air using a variety of inputs so that they can be widely used in air quality management.

4. Conclusions

Artificial neural networks are being increasingly often used in environmental research. In the research presented here, an artificial neural network (ANN) model was created to prediction of sulfur dioxides concentrations in the air of Zamość. The ANN model was created using both meterological data and those of chemical pollutants in the air, i.e.: T, RH, WS, WD, SO_2, PM10, NO_2, NO_x, CO, O_3, C_6H_6, and sines and cosines circular data. From the results, it can be concluded that ANN is characterized by 89% regression coefficient with an mean squared error (MSE) of 1.7328 for training set, 3.6460 for validation set and 4.6587 testing set at estimating air pollution in Zamość city. ANN models can be successfully got to forecast the air pollution concentrations.

In feature work we will use the advanced deep learning method to construct an intelligent model for air quality index.

References

[1] Dutta A and Jinsart W 2021 Air Pollution in Indian Cities and Comparison of MLR, ANN and CART Models for Predicting PM10 Concentrations in Guwahati, India *Asian J. Atmos. Environ.* **15** 1–26
[2] Zhang H, Liu Y, Shi R and Yao Q 2013 Evaluation of PM10 forecasting based on the artificial neural network model and intake fraction in an urban area: A case study in Taiyuan City, China *J. Air Waste Manag. Assoc.* **63** 755–63
[3] Kurnaz G and Demir A S 2022 Prediction of SO_2 and PM10 air pollutants using a deep learning-based recurrent neural network: Case of industrial city Sakarya *Urban Clim.* **41** 101051
[4] https://kobize.pl/
[5] Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe
[6] Minister Środowiska 2019 *Rozporządzenie Ministra Środowiska w sprawie programów ochrony powietrza oraz planów działań krótkoterminowych*
[7] Chu H, Wei J and Wu W 2020 Streamflow prediction using LASSO-FCM-DBN approach based on hydro-meteorological condition classification *J. Hydrol.* **580** 124253
[8] Son Y, Osornio-Vargas Á R, O'Neill M S, Hystad P, Texcalac-Sangrador J L, Ohman-Strickland P, Meng Q and Schwander S 2018 Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters *Sci. Total Environ.* **639** 40–8
[9] Xu G, Ren X, Xiong K, Li L, Bi X and Wu Q 2020 Analysis of the driving factors of PM2.5 concentration in the air: A case study of the Yangtze River Delta, China *Ecol. Indic.* **110**
[10] Fan W, Si F, Ren S, Yu C, Cui Y and Wang P 2019 Integration of continuous restricted Boltzmann machine and SVR in NOx emissions prediction of a tangential firing boiler *Chemom. Intell. Lab. Syst.* **195** 103870
[11] Murillo-Escobar J, Sepulveda-Suescun J P, Correa M A and Orrego-Metaute D 2019 Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburrá Valley, Colombia *Urban Clim.* **29** 100473
[12] Saxena A and Shekhawat S 2017 Ambient Air Quality Classification by Grey Wolf Optimizer Based Support Vector Machine *J. Environ. Public Health* **2017**
[13] Zhu S, Lian X, Wei L, Che J, Shen X, Yang L, Qiu X, Liu X, Gao W, Ren X and Li J 2018 PM2.5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors *Atmos. Environ.* **183** 20–32
[14] Kamińska J A 2018 The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław *J. Environ. Manage.* **217** 164–74
[15] Rubal and Kumar D 2018 Evolving Differential evolution method with random forest for prediction of Air Pollution *Procedia Comput. Sci.* **132** 824–33
[16] Sun H, Gui D, Yan B, Liu Y, Liao W, Zhu Y, Lu C and Zhao N 2016 Assessing the potential of random forest method for estimating solar radiation using air pollution index *Energy Convers. Manag.* **119** 121–9
[17] Wang Y, Du Y, Wang J and Li T 2019 Calibration of a low-cost PM2.5 monitor using a random forest model *Environ. Int.* **133** 105161
[18] Wang F, Yu Y, Zhao X Y, Xu J J, Xie T Y and Deresa S T 2019 Performance evaluation of reinforced recycled aggregate concrete columns under cyclic loadings *Appl. Sci.* **9**
[19] Rahman M M, Shafiullah M, Rahman S M, Khondaker A N, Amao A and Zahir M H 2020 Soft computing applications in air quality modeling: Past, present, and future *Sustain.* **12** 1–33

[20] Lu H C 2002 The statistical characters of PM10 concentration in Taiwan area *Atmos. Environ.* **36** 491–502

[21] Juahir H, Zain S M, Toriman M E, Mokhtar M and Man H C 2004 Application of artificial neural network models transportation. Langat River Basins is the most rapid urban area in Malaysia . The Langat River catchment straddles the main urban conurbation in the Klang Valley forming parts of the growing urban com **16** 42–55

[22] Perez P 2001 Prediction of maximum of 24-hour average of PM10 concentrations in Santiago, Chile *Adv. Air Pollut.* **10** 163–70

[23] Ordieres J B, Vergara E P, Capuz R S and Salazar R E 2005 Neural network prediction model for fine particulate matter (PM 2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua) *Environ. Model. Softw.* **20** 547–59

[24] Główny Inspektorat Ochrony Środowiska 2020 *Raport prognozy stężeń pyłu PM2,5 i PM10 dla lat 2020 i 2025*

[25] Bam E.K.P., Akiti T.T., Osae D.S., Ganyaglo S.Y. G A 2011 Multivariate cluster analysis of some major and trace elements distribution in an unsaturated zone profile, Densu river basin *African J. Environ. Sci. Technol.* **5** 155–67

[26] Gültepe Y and Mine A 2018 Daily SO_2 Air Pollution Prediction with the use of Artificial Neural Network Models *Int. J. Comput. Appl.* **181** 36–40

[27] Unnikrishnan R and Madhu G 2019 Comparative study on the effects of meteorological and pollutant parameters on ANN modelling for prediction of SO_2 *SN Appl. Sci.* **1** 1–12

[28] Shams S R, Jahani A, Kalantary S, Moeinaddini M and Khorasani N 2021 The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO_2 concentration *Urban Clim.* **37** 100837

[29] Chelani A B, Chalapati Rao C V., Phadke K M and Hasan M Z 2002 Prediction of sulphur dioxide concentration using artificial neural networks *Environ. Model. Softw.* **17** 159–66.