-RESEARCH PAPER-

International Journal of Mining and Geo-Engineering

IJMGE

IJMGE 56-1 (2022) 1-9

DOI: 10.22059/IJMGE.2021.313278.594874

# Determination of ozone concentration using gene expression programming algorithm (GEP)- Zrenjanin, Serbia

Hesam Dehghani [a], Milica Velicković [b], Behshad Jodeiri Shokri [c, *], Ivan Mihajlovic [b], Djordje Nikolic [b], Marija Panic [b]

[a] *Department of Mining Engineering, Hamedan University of Technology (HUT), Hamedan, Iran*

[b] *University of Belgrade, Technical Faculty in Bor, Bor, Serbia*

[c] *School of Civil Engineering and Surveying, University of Southern Queensland, Queensland, Australia*

Article History:

Received: 07 November 2020.
Revised: 04 January 2021.
Accepted: 15 January 2021.

## ABSTRACT

As one of the hazardous pollutants, ozone ($O_3$), has significant adverse effects on urban dwellers' health. Predicting the concentration of ozone in the air can be used to control and prevent unpleasant effects. In this paper, an attempt was made to find out two empirical relationships incorporating multiple linear regression (MLR) and gene expression programming (GEP) to predict the ozone concentration in the vicinity of Zrenjanin, Serbia. For this purpose, 1564 data sets were collected, each containing 18 input parameters such as concentrations of air pollutants ($SO_2$, CO, $H_2S$, NO, $NO_2$, $NO_x$, $PM_{10}$, benzene, toluene, m- and p-xylene, o-xylene, ethylbenzene), and meteorological conditions (wind direction, wind speed, air pressure, air temperature, solar radiation, and relative humidity (RH)). In contrast, the output parameter was ozone concentrate. The correlation coefficient and root mean squared error for the MLR were 0.61 and 21.28, respectively, while the values for the GEP were 0.85 and 13.52, respectively. Also, to evaluate these two methods' validity, a feed-forward artificial neural network (ANN) with an 18-10-5-1 structure has been used to predict the ozone concentration. The correlation coefficient and root mean squared error for the ANN were 0.78 and 16.07, respectively. Comparisons of these parameters revealed that the proposed model based on the GEP is more reliable and more reasonable for predicting the ozone concentrate. Also, the sensitivity analysis of the input parameters indicated that the air temperature has the most significant influence on ozone concentration variations.

**Keywords:** *Ozone concentration; Pollution; Air quality; Gene expression programming algorithm*

## 1. Introduction

In recent years, the increased concentration of atmospheric pollutants and their association with human health in metropolitan areas and developing countries has become important. Among all atmospheric pollutants, the concentration of ground-level (tropospheric) ozone produced as a result of the photochemical process has increased more than twice over the last century from 1900 to 1998 as reported by European Environment Agency (EEA) [1, 2]. Ozone re-enters the troposphere and reaches the earth's surface with the reaction with the pollutant chemicals produced and emitted at the earth's surface. In this case, ozone has a destructive and polluting role because, along with the chemicals, it severely damages other vital tissues of animals and plants. At low altitudes, ozone and smoke, and steam in the air exacerbate the pollution in many major and industrial cities worldwide. Ozone acts as greenhouse gases in the lower atmosphere (troposphere). The process of formation and depletion of the ozone layer in the lower layers is highly nonlinear [3], and the increase in the ozone density at lower points is effective in raising the earth's overall temperature. Ozone concentration is one of the major problems of air pollution for public health and the environment, which is not directly released by human activities. It is a secondary pollutant caused by the reaction of nitrous oxide ($NO_x$), carbon monoxide (CO), and volatile organic compounds (VOC) in the presence of UV rays, which are emitted by human activities [4- 7]. The high oxidizing potential of ozone during breathing can cause serious respiratory problems, chest pain, asthma attacks, bronchitis, headaches, and many other human problems. Therefore, several different but related models need to be developed. To predict future ozone concentration, it is necessary to develop a model that can describe the complex relationships between the ground-level ozone and a large number of variables involved in its formation or depletion. There are various methods in the scientific literature for predicting ozone concentration. The predictions are based on the development of three different models [8]: (a) deterministic models, (b) physical models, and (c) statistical models. The atmospheric scattering models simulating the atmosphere's physical and chemical processes have been previously used to predict the ozone concentration [9]. However, such models are not suitable because they require the knowledge of specific aspects of the reaction, complex calculations, and a large number of input parameters [3, 10]. Keeping these drawbacks in mind, the statistical models have begun to be used as alternatives to deterministic and physical models for the ozone level prediction. The simplicity of the models shows their advantage; however, a serious disadvantage is a fact that they only relate to a particular area because they are based on local data [8, 11, 12]. Although the linear models are easier to use and widely accepted, they do not consider the ozone's nonlinearities. On the other hand, these defects can be remedied by neural networks [13]. For many years, researchers have estimated the ozone concentration using various smart methods. Arsić et al. investigated the prediction of ozone concentration in the ambient air using the linear multivariate regression

* Corresponding author. Tel: +61416080049, Fax: +61746312920, *E-mail address:* behshad.jodeirishokri@usq.edu.au (B. Jodeiri Shokri).

and artificial neural network methods. They used the statistical modeling results of ground-level ozone concentration in the vicinity of Zrenjanin, Serbia. They considered two categories of parameters in their study: concentration of $SO_2$, CO, $H_2S$, NO, $NO_2$, $NO_x$, $PM_{10}$, benzene, toluene, m- and p-xylene, o-xylene, ethylbenzene in air and meteorological parameters (wind direction, wind speed, air pressure, air temperature, solar radiation and relative humidity (RH)) as input parameter and ozone concentration as an output parameter. The multiple linear regression (MLR) and artificial neural network (ANN) analyses were used as a tool for the mathematical analysis of the occurrence. The results showed that ANN provides a better estimate of ozone concentration, while the MLR model once again proved to be less efficient in accurately predicting the ozone concentration [14]. Mishra and Goyal investigated the neuro-fuzzy approach to predict the ozone episodes in India's Delhi metropolitan area. The MLR, ANN, and neuro-fuzzy (NF) artificial intelligence techniques were applied. The air pollutants and meteorological parameters were utilized to analyze the ozone episode. Also, correlation coefficient (R), normalized mean square error (NMSE), fractional bias (FB), and index of agreement (IOA) was considered as the objective functions. The statistical analysis showed that artificial intelligence implementation has a more reasonable agreement with the observed values [4]. Fontes et al. addressed the question as to whether artificial neural networks can be used to predict the source of ozone episodes. They used the multilayer perceptron (MLP) with a hidden layer to automate the ozone episode classification according to air quality and meteorological variables from long data series between 2001 and 2010. They found that with a small complexity model, the mean error is obtained about 2-7% (depending on the scenario), which can be a good generalization. The results suggested that such a tool can be used to help the authorities manage the ozone [15]. Samdian-Fard et al. made a comparative analysis of ozone level prediction models using gene expression programming (GEP) and the MLR. In this study, they presented the results of two diagnostic models including GEP, which is a variable of genetic programming (GP) and MLR to predict the ozone levels in real-time up to 6 hours ahead in four stations in Bilbao, Spain. They considered the GEP inputs as weather conditions (wind speed and direction, temperature, relative humidity, pressure, solar radiation, and thermal gradient), hourly ozone level, and traffic parameters (number of vehicles, occupation rate, and speed), which were measured from 1993 to 1994. They compared the performance of the developed models with the observed values and evaluated the model using specific performance measures for the air quality models developed in the validation model and recommended by the US Environmental Protection Agency (USEPA). They found that GEP provides superior predictions in most cases [16]. Baawain and Al-Serihi used a systematic approach to predict air pollution at ground level (around an industrial port) using ANN. They proposed a precise method of preparing the air quality data for achieving the more accurate air pollution prediction models based on the ANN. The models predict the daily concentrations of air pollutants at ground level, namely CO, $PM_{10}$, NO, $NO_2$, $NO_x$, $SO_2$, $H_2S$, and $O_3$ measured by the air quality control station in Ghadafan village. The models' training is based on the MLP method with the back-propagation (BP) algorithm. The results indicated an excellent agreement between the actual and predicted concentrations since the values of the multiple coefficients of determination ($R^2$) for all ANN models exceed 0.70. The results also revealed that the importance of temperature in the daily variations of $O_3$, $SO_2$, and $NO_x$. The wind speed and wind direction play important roles in the daily variations of NO, CO, $NO_2$, and $H_2S$. The $PM_{10}$ concentration was affected by almost all the measured meteorological parameters [17]. Geng et al. measured the properties of ozone, $NO_x$, and VOC in Shanghai, China. They investigated the spatial and temporal variations of $O_3$ and its precursors ($NO_x$ and VOCs) and the precursors' influence on $O_3$ formation. A chemical mechanism model (NCAR-MM) was used to evaluate the sensitivity of $O_3$ formation to $NO_x$ and VOC concentrations. The results show that the concentration of $O_3$ in rural areas is higher than that in the city center. The results also indicated that the highly reactive (aromatic) VOCs play an important role in determining the $O_3$ formation. The toluene/benzene ratio showed that cars play an important role in forming $O_3$ in Shanghai. The further increase in Shanghai vehicles could lead to high potential $O_3$ concentration in the future [18]. Stathopoulou et al. investigated the effect of temperature on the tropospheric ozone ($O_3$) concentration in Athens' urban and photochemical polluted areas. The hourly values of ambient air temperature used to study the impact of urban heat islands were recorded in twenty-three experimental stations in Athens, while the ozone concentration was measured at three stations for two years (1996 to 1997). The linear correlation between ozone concentration and air temperature and the temporary changes of temperature and ozone concentrations were calculated and analyzed for the test stations. Besides, they used a neural network method to investigate the influence of temperature on ozone concentration values in Athens's greater region. The neural network model used ambient air temperature as one of the input parameters and found that temperature is a dominant parameter that significantly affects the ozone concentration values [19]. Sousa et al. used MLR and principal component-based artificial neural networks to predict the ozone concentration. They compared the developed model with MLR, ANN-based on principal data, and principal component regression. The results showed that the use of principal components as input improves the model prediction by reducing the complexity and eliminating data [20]. Bandyopadhyay and Chattopadhyay investigated the single hidden-layer ANNs models compared with MLR models to predict the total ozone time series in Arosa, Switzerland. The single-layer neural network model was developed with a variable number of nodes. The performance was evaluated by the least-squares method and error estimation, and compared with MLR models. Consequently, they identified a hidden layer model with 8 hidden nodes as the best prediction model [21]. Wang et al. investigated daily maximum ozone level prediction using the combined neural networks and statistical properties. This study aimed to develop an improved neural network model that combines adaptive radial basis function (ARBF) of the network with the statistical properties of ozone in the selected specific regions and is used to predict the maximum daily ozone level. The improved method uses the hourly time series data collected at three air pollutant monitoring stations in Hong Kong during 1999 and 2000. The simulation results show the effectiveness and reliability of the proposed method [22]. Nishanth et al. studied the changes in the ground-level ozone and $NO_x$ concentration in India. The results showed that the highest concentrations of ozone in the afternoon and the minimum values in the early hours of the morning and reported the highest ozone mix ratio values in winter [23]. Prybutok et al. developed a neural network model to predict the daily maximum ozone levels and compared it with two conventional statistical models, regression and Box-Jenkins ARIMA [24]. Zhu et al. developed a novel ozone prediction model based on the ozone generation mechanism in the corona discharge region [25]. The results showed that the neural network model is more accurate than the tested regression and Box-Jenkins ARIMA models. The ground-level ozone has been measured in Serbia since 2008. However, few comprehensive studies have been conducted on the formation, concentration, and potential risks for human health. For this reason, this study aimed to determine the probable pattern of dependence, on the one hand, between the concentration of ozone in the ambient air and, on the other hand, other pollutants and meteorological conditions. Such a model makes it possible to develop other measures that can be used to control the concentration of ozone in the ambient air. The Banat region (northeast of Serbia) represents Europe's most fertile agricultural areas as the measurement site. The increased ozone concentration has been recorded in this area, which is a potential risk to human health and plant growth [26]. Hybrid models, combining the benefits of some models, are suggested to achieve better prediction, and the decomposition approaches such as CEEMD enhance the performance of hybrid models [27]. Alomar et al. were applied the wavelet transform (WT) approach to handling input variables before introducing them to ANN. This approach attempts to remove the noise impacts, which decreases the accuracy of simulation processes. Additionally, to improve ANN model performance, selecting a suitable type of transfer function with effective input combinations was thoroughly investigated before introducing the

WT approach. The hybrid model (W-ANN) was also compared to the classical ANN in predicting one-hour ozone concentrations. The results show that the hybrid model (W-ANN) reported fewer errors than the conventional ANN modeling approach [28]. According to the authors' best review, many AI techniques were developed and proposed for predicting future ozone concentrations. However, their effectiveness is different. Furthermore, depending on the meteorological condition, amount of pollutant, and the location of each measurement station, the ozone, and its effects are different. The previous studies' biggest advantage was to consider the effect of different parameters on the ozone concentration. Nevertheless, few studies have attempted to minimize this phenomenon in addition to predicting it. Considering the disadvantages caused by increasing the ozone amount and its importance from many researchers' viewpoints, this paper studied the undesirable ozone concentration and its influential parameters in Zrenjanin, Serbia. For achieving this aim, a mathematical equation was developed using GEP.

## 2. Gene expression programming

Gene expression programming (GEP) algorithm is a method for developing computer programs and mathematical modeling based on evolutionary computation inspired by natural evolution. This method was devised by Ferreira in 1999 and formally introduced in 2001 [29]. Since then, GEP has been applied in different scopes of science as reliable predicting tools. For instance, Jodeiri Shokri et al. (2020), used GEP for estimating silver price by using historical data [30]. In another research, Jodeiri Shokri et al. (2020) applied this algorithm for investigating acid mine drainage (AMD) throughout copper tailing particles [31]. Shakeri et al. (2020) predicted blast-induced ground vibration by coupling GEP and MLR [32].

The GEP algorithm combines the two former inheritance algorithms' dominant view to cover both ones' weaknesses. In this method, the chromosome genotype has a linear structure similar to that of the genetic algorithm (GA). The chromosomes' phenotype is a tree structure with variable length and size similar to the genetic programming (GP) algorithm. The GEP algorithm employs the Karva language. The multiple genes are used for multiple chromosome structures and create sub-trees from multiple genes that provide better adaptability and responsiveness to the algorithm. The flowchart or process of the GEP algorithm is shown in Figure 1, where the initial stage of the algorithm of random population generation (chromosome) begins. The generated population is then expressed, and each individual is evaluated and selected according to the evaluation function. By modifying the selected individuals, they begin to replicate the populations with new traits. Like the older generation, the new individuals repeat the processes that continue until the right solution is reached [29].

The simplicity of the gene expression programming algorithm allows encoding any part of the program, making the evolution effective. The expression tree and chromosome are the two main parameters of this algorithm, where the expression trees display the information encoded on the chromosomes. Displaying this information is done by converting the information from the chromosome to the expression tree. This code is a one-to-one relationship between chromosomes, functions, and terminals. The linear chromosome components include the terminals (A, B, C, ...) and functions (+, -, ...). In the GEP algorithm, the lengths of chromosomes and genes are constant, and only the length of the open reading frame (ORF) changes. This causes the GEP terminal not to match the terminal of the genes. This matching is because of the non-coding regions at the end of the gene. The non-coding regions in GEP allow the operators to work without restriction and generate genetic diversity to achieve evolution. According to Equation (1), genes are composed of two parts, Head (h) and Tail (t), each having different functions. The head part is used to encode the functions, and a tail part is a place for the terminals to ensure the formation of a good structure. The number of function arguments means the number of variables the function needs for its operation. For example, the sin (x) function has one argument, and the if (x, y, z) function has three arguments. There are the following rules for drawing an expression tree, and a hypothetical example in Figure 2 is used for a better understanding:

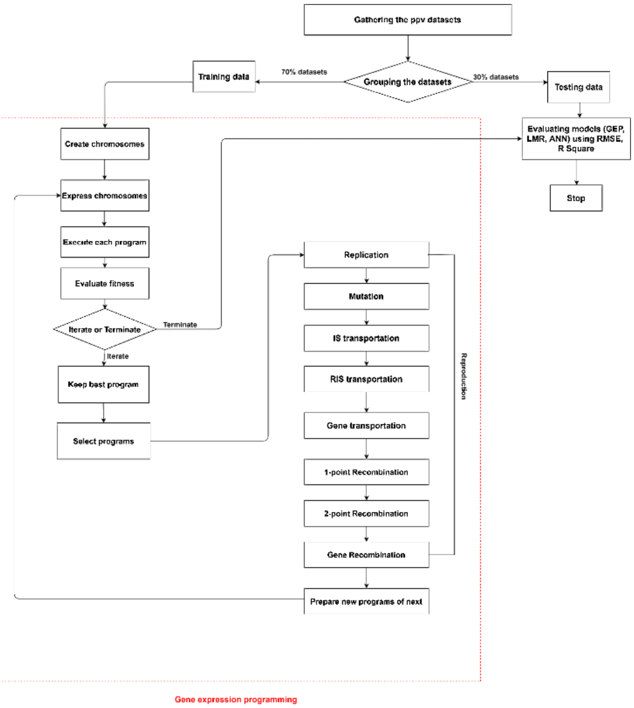


**Figure 1.** Flowchart of GEP algorithm

1. Read the root of the expression tree, which lies above the tree.
2. Specify the number of output nodes depending on the root (for example, the root has one output node).
3. Provide existing functions after the root and define output nodes.
4. Make the number of nodes in the next row equal to the current row arguments' sum.
5. Fill the nodes in each row from left to right in the same order as the gene members.
6. Continue this process until having the only terminal at the end [33].

$$t = h\ (n-1) + 1 \tag{1}$$

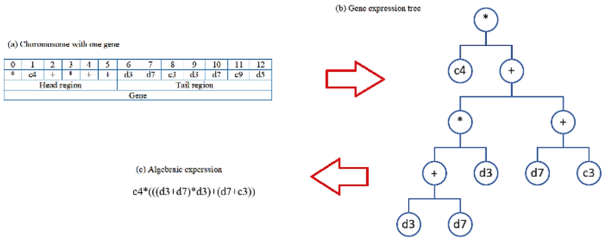


**Figure 2.** (a) Karva language, (b) expression tree, (c) final relation

## 3. Case study

The present research was conducted in the Banat region (Serbia), one of the most fertile parts of Europe. This region extends over three countries: the northern part belongs to Hungary, the eastern part belongs to Romania, and the western part belongs to Serbia. The Serbian part of Banat covers an area of 8997 km2. Zrenjanin is the center of this region, with a population of 80,000. The climate, typical for this area, is moderate continental with four seasons, while the average annual temperature is 11.2°C. Wind direction is mostly east, southeast or northwest. The average number of sunny hours in the area is 2000 to 2200 per year, and the average relative air humidity is 75%. A detailed description of the study area can be found in [14, 26].

Continuous measurement of the air pollutants, used for the modeling of Ozone concentration presented in this research, was facilitated in an automatic measuring station located in Zrenjanin. This station measures air pollution levels originating from exhaust gasses and other sources of pollution. Details on the measurement intervals, calibration of the equipment, quality control, and standardization are presented in [26].

## 4. Data collection

To model the ozone concentration, the required data were used from an automated measurement station in the vicinity of Zrenjanin (Serbia). The data were collected over one year at different times so that all pollutants and all meteorological parameters were simultaneously measured. This period is as follows:

1. Winter: February 1-8 and 23-28; December 21-31;
2. Spring: May 5-15;
3. Summer: July 13-18; September 1-20; and
4. Fall: October 1-22.

During these 83 days, the measurements were made hourly, from 0:00 to 24:00, and the hourly averages were calculated. As such, a representative database was created for accurate statistical analysis consisting of 1564 data series. The obtained database is shown in Table 1. After the investigations, 18 data were used as input parameters and the ozone concentration as an output parameter to construct the ozone concentration, prediction model. The statistical indicators of the input and output parameters are presented in Table 1.

**Table 1.** Statistical indicators of input and output parameters

| Type | Parameter | Symbol | Range | Min | Max | Mean | Std. Deviation | Variance |
|---|---|---|---|---|---|---|---|---|
| Input | $SO_2$ (µg/$m^3$) | d0 | 220.40 | 0 | 220.4 | 17.74 | 22.59 | 510.09 |
| | CO (µg/$m^3$) | d1 | 480.00 | 0.7 | 480.7 | 31.09 | 37.52 | 1407.77 |
| | $H_2S$ (µg/$m^3$) | d2 | 150.30 | 4 | 154.3 | 34.07 | 22.79 | 519.46 |
| | NO (µg/$m^3$) | d3 | 864.40 | 5.5 | 869.9 | 81.56 | 74.91 | 5611.21 |
| | $NO_2$ (µg/$m^3$) | d4 | 6337.00 | 0 | 6337 | 781.27 | 628.44 | 394931.80 |
| | $NO_x$ (µg/$m^3$) | d5 | 73.91 | 0 | 73.91 | 2.01 | 6.19 | 38.36 |
| | $PM_{10}$ (µg/$m^3$) | d6 | 497.40 | 0 | 497.4 | 43.85 | 38.75 | 1501.18 |
| | Benzene (µg/$m^3$) | d7 | 22.23 | 0 | 22.23 | 1.78 | 2.62 | 6.87 |
| | Toluene (µg/$m^3$) | d8 | 90.40 | 0 | 90.4 | 2.71 | 4.24 | 17.99 |
| | m,p-Xylene (µg/$m^3$) | d9 | 34.14 | 0 | 34.14 | 1.66 | 2.87 | 8.24 |
| | o-Xylene (µg/$m^3$) | d10 | 25.81 | 0 | 25.81 | 0.62 | 1.76 | 3.10 |
| | Ethylbenzene (µg/$m^3$) | d11 | 15.97 | 0 | 15.97 | 0.53 | 1.45 | 2.10 |
| | Wind direction (º) | d12 | 345.00 | 9 | 354 | 188.69 | 70.58 | 4981.78 |
| | Wind speed (m/s) | d13 | 452.82 | 0.18 | 453 | 2.03 | 11.56 | 133.67 |
| | Air temperature (ºC) | d14 | 47.70 | -12.5 | 35.2 | 15.11 | 9.70 | 94.00 |
| | Air pressure (hPa) | d15 | 993.00 | 27 | 1020 | 993.15 | 113.24 | 12823.18 |
| | Solar radiation (W/$m^2$) | d16 | 844.00 | 4 | 848 | 139.87 | 215.66 | 46510.81 |
| | Relative humidity-RH (%) | d17 | 82.00 | 10 | 92 | 64.71 | 16.98 | 288.46 |
| Output | Ozone concentration (g/$m^3$) | y | 160.70 | 1.3 | 162 | 69.23 | 34.35 | 1179.70 |

In the next step, the data were divided into two parts: model-building data and validation data, where 70% (1036 data) were used for the model building, and 30% (528 data) were used for the validation. The training and validation data were randomly selected.

## 5. Modeling of ozone concentration using GEP

In the present study, GEP software was used to analyze the final relationship between the initial and ozone concentrations. GEP modeling consists of five main steps:

***Step I:*** The first step in the GEP modeling is to select the fitness function. In this study, the MSE, RMSE, and RRSE fitness functions were used to predict the ozone concentration. The relationships between each of the fitness functions are listed below. Among these functions, based on the results in Table 2, the RRSE fitness function yielded the best results for predicting the ozone concentration.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{Xipred} - \mathrm{Ximes})^2 \qquad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{Xipred} - \mathrm{Ximes})^2} \qquad (3)$$

$$\mathrm{RRSE} = \sqrt{\frac{\sum_{i=1}^{N}(\mathrm{Xipred} - \mathrm{Ximes})^2}{\sum_{i=1}^{N}(\mathrm{Ximes} - \mathrm{Ximes})^2}} \qquad (4)$$

Where, $X_{ipred}$ is the predicted Ozone concentration, and $X_{imes}$ is the measured Ozone concentration.

***Step II:*** This step involves selecting a set of terminals and functions to form the chromosomes. The terminals are selected according to the inputs and outputs of the model. In the present study, 18 parameters listed in Table 1 are considered input and the ozone concentration. Also, at this step, the most appropriate functions are determined to obtain the final Equation. The functions used in the model include {F = +, -, *, /, sqrt, x, x^2, x^3, x^(1/3), 1/x, exp(x), log(x)}.

***Step III:*** This step involves selecting the chromosomal structure. This step is done by trial and error. For this purpose, more than 40 models were developed. In this paper, 30 chromosomes were selected.

***Step IV:*** In this step, the type of link function is select. In this paper, the sum link function was used to predict the ozone concentration in all models.

***Step V:*** Finally, a set of genetic and rate operators are produced at this step. Table 3 lists the parameters used in this software [30]. Finally, the values of $R^2$ and other parameters were obtained concerning the models in Table 2, which had the best results among all models, and the results are listed in Table 4.

**Table 2.** Values of parameters used to predict ozone concentration

| GEP Parameter | Model | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Fitness function | RMSE | RMSE | MSE | RRSE | RRSE |
| Inversion rate | 0.00546 | 0.00546 | 0.00546 | 0.00546 | 0.00546 |
| IS transportation rate | 0.00546 | 0.00546 | 0.00546 | 0.00546 | 0.00546 |
| RIS transportation rate | 0.00546 | 0.00546 | 0.00546 | 0.00546 | 0.00546 |
| One-point recombination rate | 0.00277 | 0.00277 | 0.00277 | 0.00277 | 0.00277 |
| Two-point recombination rate | 0.00277 | 0.00277 | 0.00277 | 0.00277 | 0.00277 |
| Gene size | 15 | 21 | 21 | 19 | 17 |
| Head size | 7 | 10 | 10 | 9 | 8 |
| Tail size | 8 | 11 | 11 | 10 | 9 |
| Mutation rate | 0.00138 | 0.00138 | 0.00138 | 0.00138 | 0.00138 |
| Number of Chromosome | 30 | 30 | 30 | 30 | 30 |
| Number of genes | 4 | 3 | 4 | 3 | 4 |
| Gene recombination rate | 0.00277 | 0.00277 | 0.00277 | 0.00277 | 0.00277 |
| Gene transportation rate | 0.00277 | 0.00277 | 0.00277 | 0.00277 | 0.00277 |
| Training | 70% | 70% | 70% | 70% | 70% |
| Validation | 30% | 30% | 30% | 30% | 30% |
| Number of generation | 10000 | 10000 | 10000 | 10000 | 10000 |

**Table 3.** GEP functions definition

| Name | Representation | Definition |
|---|---|---|
| Addition | + | (x+y) |
| Subtraction | - | (x-y) |
| Multiplication | * | (x*y) |
| Division | / | (x/y) |
| Exponential | Exp | exp(x) |
| Natural logarithm | Ln | ln(x) |
| Inverse | Inv | 1/x |
| x to the power of 2 | X2 | x^2 |
| x to the power of 3 | X3 | x^3 |
| Cube root | 3Rt | x^(1/3) |
| Addition with 3 inputs | Add3 | (x+y+z) |
| Multiplication with 3 inputs | Mul3 | (x*y*z) |
| Average of 2 inputs | Avg2 | avg(x,y) |
| Sine | Sin | sin(x) |
| Cosine | Cos | cos(x) |
| Tangent | Tan | tan(x) |
| Arctangent | Atan | arctan(x) |
| Hyperbolic tangent | Tanh | tanh(x) |
| Complement | NOT | (1-x) |

**Table 4.** Performance indicator values for building GEP models to predict ozone concentration

| Model | Fitness function | Training | | Testing | |
|---|---|---|---|---|---|
| | | $R^2$ | RMSE | $R^2$ | RMSE |
| 1 | RMSE | 0.76 | 16.77 | 0.75 | 17.09 |
| 2 | RMSE | 0.73 | 17.81 | 0.73 | 17.96 |
| 3 | MSE | 0.75 | 17.09 | 0.75 | 16.96 |
| 4 | RRSE | 0.79 | 16.44 | 0.85 | 13.52 |
| 5 | RRSE | 0.76 | 16.82 | 0.76 | 16.78 |

According to Table 4, among the models built using the GEP, Model 4 was selected as the best model among the five final models due to the high $R^2$ value and low RMSE value. Using the GEP algorithm, the last relationship was obtained to predict the ozone concentration in the expression tree (Figure 3) and the last relationship (Equation 5). In this Equation, $C_0$ to $C_9$ are the numerical constants.

$$\boldsymbol{y} = \left(\boldsymbol{c_9}^2 - \left(\sqrt[3]{(\boldsymbol{d_2} - \boldsymbol{d_{13}}) - (\boldsymbol{d_{14}} \times (\boldsymbol{d_{14}} \times \boldsymbol{c_4}))}\right)\right) + \left(\left(\left(\boldsymbol{d_{14}} \times \sqrt{\boldsymbol{d_8} \times \boldsymbol{c_2}}\right) - \boldsymbol{d_8}\right) - \boldsymbol{c_2}\right) + \left(\boldsymbol{d_{14}} - \left(\left(\sqrt[3]{\sqrt[3]{\boldsymbol{d_8}^3 \times (\boldsymbol{d_3} \times \boldsymbol{d_{10}})}}\right) \times \boldsymbol{d_{14}}\right)\right) \quad (5)$$

The values of the final constant coefficients in Equation (5) are given in Table 5. By substituting the coefficients into Equation (5), the final Equation (6) is obtained

**Table 5.** Constant coefficients in the GEP model

| $C_0$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ | $C_9$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 2.37 | 0.0 | 4.93 | 0.0 | 0.0 | 0.0 | 0.0 | 6.51 |

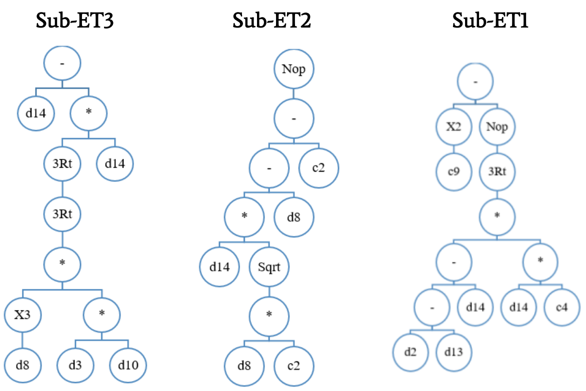


**Figure 3.** Expression tree for GEP model to predict ozone concentration

.

$$\mathbf{y} = \left((6.51)^2 - \left(\sqrt[3]{(H_2S - wind\ speed) - (Air\ temprature \times (Air\ temorature \times 4.93))}\right)\right) + \left(\left(\left(Air\ temprature \times \sqrt{Toluene \times 2.37}\right) - Toluene\right) - 2.37\right) + \left(Air\ temprature - \left(\left(\sqrt[3]{\sqrt[3]{Air\ temprature^3 \times (NO \times oXylene)}}\right) \times Air\ temprature\right)\right) \quad (6)$$

## 6. Validation

A new mathematical model was developed in the present paper using the gene expression programming algorithm to predict the ozone concentration in ambient air with effective parameters. The obtained results can be compared with the models obtained from other methods to evaluate the proposed relation's efficiency. The multivariate regression and artificial neural networks were used for this purpose. In the following, the models proposed by these methods are presented, and finally, the results are compared with the determination coefficient ($R^2$), and root mean squared error (RMSE) indicators.

### *6.1. Multivariate regression*

The linear dependence of the air ozone concentration (Y) on the input parameters (d0-d18) was obtained using IBM SPSS v25 software. Among the constructed models, the results of the top four models are shown in Table 6. It should be noted that the statistical analyses have been extensively used in suggesting empirical relationships in various scope of science. For instance, Jodeiri Shokri et al. (2014) suggested several relationships for predicting AMD generation throughout coal waste particles [34]. Soleimani and Jodeiri Shokri applied statistical methods for predicting chromite ore production rates in Iran [35]. Comparing the obtained results shows that model 4 with an R-square of 0.63 and RMSE of 21.0781 was selected as the best model for the training data. The best MLR model is presented in Eq. 7. According to Equation 7, the R-square and RMSE values for the validation data were 0.61 and 21.28, respectively.

the Figure 4 demonstrates a histogram for the analysis of the modeling error. The modeling error distribution function is normal, confirming regression test has been done correctly.

$$\boldsymbol{y} = 87.948 + (1.210 \times Air\ temprature) + (-27.343 \times Ethylbenzene) + (12.319 \times mpXylene) + (-0.007 \times NO_2) + (-0.439 \times RH) + (0.156 \times PM_{10}) + (-0.794 \times NO_x) + (-0.201 \times H_2S) + (-3.940 \times Wind\ speed) + (0.099 \times SO_2) + (-1.910 \times oXylene) + (0.012 \times Solar\ radiation) + (-1.491 \times Benzene) \quad (7)$$

**Table 6.** Comparison of statistical criteria of models obtained using MLR for training data

| No. | Model | R Square | RMSE |
|---|---|---|---|
| 1 | $\boldsymbol{y} = 89.403 + (1.438 \times d_{14}) + (-28.414 \times d_{11}) + (10.729 \times d_9) + (-0.005 \times d_4) + (-0.497 \times d_{17}) + (0.134 \times d_6) + (-0.827 \times d_5) + (-0.249 \times d_2) + (-3.399 \times d_{13}) + (0.114 \times d_0) + (-1.910 \times d_{10}) + (0.012 \times d_{16}) + (-1.491 \times d_7)$ | 0.624 | 21.2315 |
| 2 | $\boldsymbol{y} = 89.345 + (1.431 \times d_{14}) + (-26.109 \times d_{11}) + (10.809 \times d_9) + (-0.005 \times d_4) + (-0.495 \times d_{17}) + (0.137 \times d_6) + (-0.816 \times d_5) + (-0.246 \times d_2) + (-3.382 \times d_{13}) + (0.109 \times d_0) + (-2.161 \times d_{10})$ | 0.626 | 21.1748 |
| 3 | $\boldsymbol{y} = 84.541 + (1.362 \times d_{14}) + (-27.246 \times d_{11}) + (11.376 \times d_9) + (-0.006 \times d_4) + (-0.428 \times d_{17}) + (0.136 \times d_6) + (-0.813 \times d_5) + (-0.215 \times d_2) + (-3.839 \times d_{13}) + (0.108 \times d_0) + (-2.200 \times d_{10}) + (0.010 \times d_{16})$ | 0.628 | 21.1254 |
| 4 | $\boldsymbol{y} = 87.948 + (1.210 \times d_{14}) + (-27.343 \times d_{11}) + (12.319 \times d_9) + (-0.007 \times d_4) + (-0.439 \times d_{17}) + (0.156 \times d_6) + (-0.794 \times d_5) + (-0.201 \times d_2) + (-3.940 \times d_{13}) + (0.099 \times d_0) + (-1.910 \times d_{10}) + (0.012 \times d_{16}) + (-1.491 \times d_7)$ | 0.630 | 21.0781 |

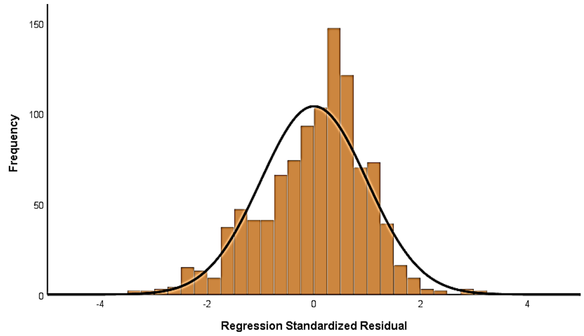


**Figure 4.** Histogram for the analysis of the modeling error

### *6.2. Artificial Neural Network*

For finding the ability of GEP in predicting the ozone concentration, the obtained results were compared with the results of the artificial neural network algorithm. This algorithm is so accurate, and many researchers used it for predicting various problems [36, 37]. For this purpose, a multilayer perceptron (MLP) network was used in MATLAB software. The training and validation datasets were selected randomly. To reach an appropriate architecture, MLP networks with one and two hidden layers were examined. To determine the optimum network, RMSE was calculated for various models. The network with architecture 18-20–5–1 (LOGSIG–LOGSIG-LOGSIG–POSLIN), which has the minimum RMSE, is considered as the optimum model (Figure 5).

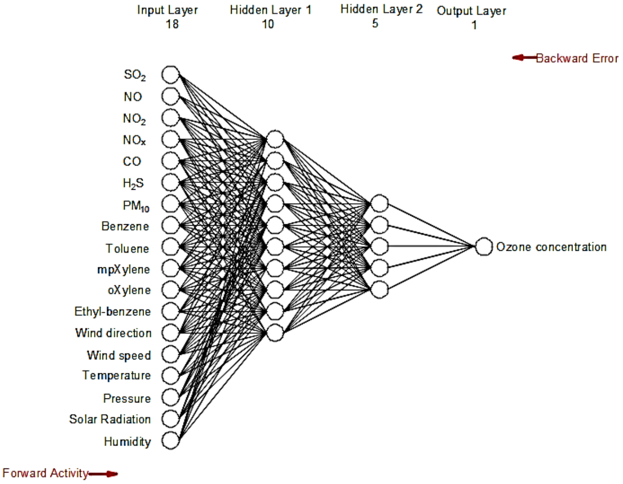
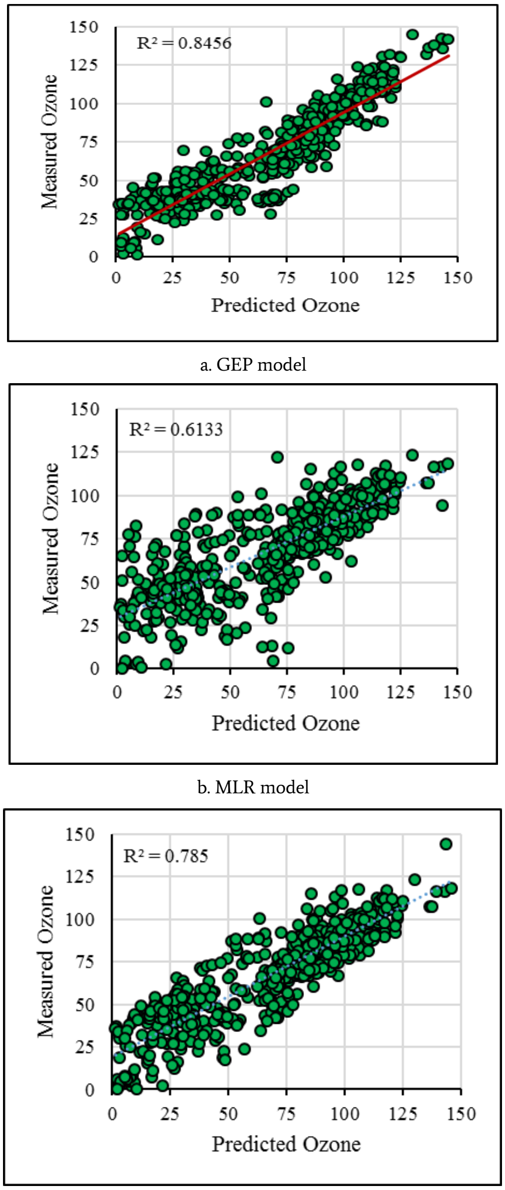


**Figure 5.** the ANNs architecture with 18-10-5-1 (back-propagation network)

Figure 6 compares the validation R-square of GEP, MLR, and ANNs models. Based on this figure, the best predictions were obtained using the GEP algorithm. The amount of absolute relative error (ARE), average absolute relative error (AARE), and root mean square error (RMSE), were compared for GEP, MLR, and ANNs in Figure 7.

## 7. Sensitivity Analysis

A useful concept has been proposed to identify the significance of each "cause" factor (the input data) on the "effect' factor" (the output). This enables the most sensitive factors affecting ozone concentration to be identified hierarchically. To find the most influential parameters in ozone concentration, a sensitivity analysis was performed on the input parameters. For achieving this aim, two types of sensitivity analysis i.e., tornado and spider graphs were conducted. These plots are typically created by fixing an input distribution to a low value (say its fifth percentile), running a simulation, recording the output means, and then repeating the process with a high value (say 95th percentile) of the input distribution: these output means to define the extremes of the bars. The input parameters were chosen based on the result of the GEP model.



a. GEP model

b. MLR model

C. ANN model

**Figure 6.** Comparison of the measured Ozone versus predicted PPV using the GEP, ANNs, and MLR for test data

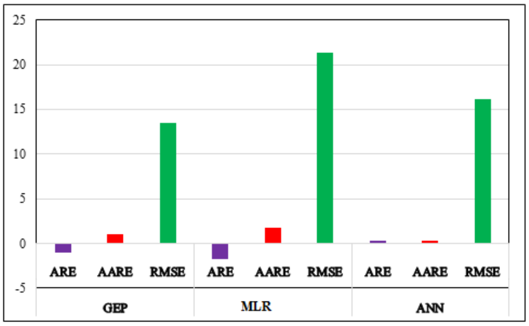


**Figure 7.** Comparison of the statistical parameters for predictions of MLR, ANNs, and GEP models

The results of the sensitivity analysis were shown in Figures 8 and 9. In the tornado sensitivity analysis, the ranges of correlations are between -1 and +1. Figure. 8 shows tornado analysis for ozone concentration. As it is shown in this figure, air temperature and Toluene amount are the most influential parameters on ozone concentration. The ozone concentration variation based on changing the air temperature, toluene, o-Xylene, $H_2S$, and NO is shown in Figure 9. As mentioned in this figure, changes in "air temperature" affect Ozone concentration. It is evident that with increasing the amount of air temperature, Ozone concentration increases dramatically.

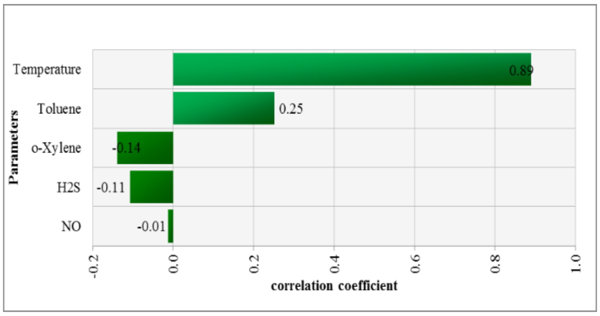


**Figure 8.** Tornado analysis

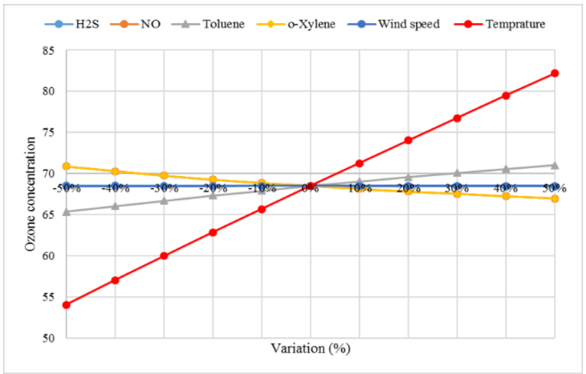


**Figure 9.** Spider analysis

## 8. Conclusions

The present study predicted the concentration of ozone in ambient air using GEP, MLR, and ANNs algorithms. For this purpose, 1564 data sets were collected each containing 18 input parameters such as concentrations of air pollutants ($SO_2$, CO, $H_2S$, NO, $NO_2$, $NO_x$, $PM_{10}$, benzene, toluene, m- and p-xylene, o-xylene, ethylbenzene), and meteorological conditions (wind direction, wind speed, air pressure, air temperature, solar radiation, and relative humidity (RH)), in the vicinity of Zrenjanin (Serbia). Once finished with developing various models for predicting the ozone concentration, some performance indicators were calculated to evaluate the proposed prediction models, including R-square, RMSE, and RRSE. The results showed that the developed GEP could practically outperform the MLR. Taking RRSE as the objective function, the obtained values of R-square and RMSE using the GEP algorithm for predicting the ozone concentrate indicated higher accuracy of this algorithm compared to the MLR and the ANNs. The obtained R-squared values and RMSE corresponding to GEP were 0.85 and 13.52, respectively, while those of MLR were 0.61 and 21.28, respectively. Also, the lower values of R-Squared and RMSE, 0.79 and 16.07, for the ANNs results proved that the GEP was more reliable and more reasonable. Eventually, the results' sensitivity analysis showed that the air temperature has the most significant effect on the increased ozone concentration. Therefore, global warming can help increase the concentration of this pollutant more rapidly in the earth's environment.

## REFERENCES

[1]. USEPA, "Ozone Pollution", www.epa.gov/ozone-pollution, 2017.

[2]. Susaya, J., Kim, K. H., Shon, Z. H., & Brown, R. J. (2013). Demonstration of long-term increases in tropospheric O3 levels: Causes and potential impacts. Chemosphere, 92(11), 1520-1528.

[3]. Abdul-Wahab, S. A., & Al-Alawi, S. M. (2002). Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environmental Modelling & Software, 17(3), 219-228.

[4]. Mishra, D., & Goyal, P. (2016). Neuro-Fuzzy Approach to forecasting Ozone Episodes over the urban area of Delhi, India. Environmental Technology & Innovation, 5, 83-94.

[5]. Duan, J., Tan, J., Yang, L., Wu, S., & Hao, J. (2008). Concentration, sources, and ozone formation potential of volatile organic compounds (VOCs) during ozone episode in Beijing. Atmospheric Research, 88(1), 25-35.

[6]. Lengyel, A., Héberger, K., Paksy, L., Bánhidi, O., & Rajkó, R. (2004). Prediction of ozone concentration in ambient air using multivariate methods. Chemosphere, 57(8), 889-896.

[7]. Shao, M., Zhang, Y., Zeng, L., Tang, X., Zhang, J., Zhong, L., & Wang, B. (2009). Ground-level ozone in the Pearl River Delta and the roles of VOC and NOx in its production. Journal of Environmental Management, 90(1), 512-518.

[8]. Sharma, S., Sharma, P., & Khare, M. (2017). Photo-chemical transport modelling of tropospheric ozone: A review. Atmospheric Environment, 159, 34-54.

[9]. Moussiopoulos, N., Sahm, P., & Kessler, C. (1995). Numerical simulation of photochemical smog formation in Athens, Greece—a case study. Atmospheric Environment, 29(24), 3619-3632.

[10]. Chaloulakou, A., Saisana, M., & Spyrellis, N. (2003). Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Science of the Total Environment, 313(1-3), 1-13.

[11]. Sousa, S. I. V., Martins, F. G., Alvim-Ferraz, M. C. M., & Pereira, M. C. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environmental Modelling & Software, 22(1), 97-103.

[12]. Abdi-Oskouei, M., Carmichael, G., Christiansen, M., Ferrada, G., Roozitalab, B., Sobhani, N., Wade, K., Czarnetzki, A., Pierce, R.B., Wagner, T., & Stanier, C. (2020). Sensitivity of meteorological skill to the selection of WRF-Chem physical parameterizations and impact on ozone prediction during the Lake Michigan Ozone Study (LMOS). Journal of Geophysical Research: Atmospheres, 125(5), e2019JD031971.

[13]. Đorđević, P., Mihajlović, I., & Živković, Ž. (2010). Comparison of linear and nonlinear statistics methods applied in the industrial process modeling procedure. Serbian Journal of Management, 5(2), 189-198.

[14]. Arsic, M., Nikolic, D. J., Mihajlovic, I., & Zivkovic, Z. (2014). Monitoring of surface ozone concentrations in the western Banat region (Serbia). Applied ecology and environmental research, 12(4), 975-989.

[15]. Fontes, T., Silva, L. M., Silva, M. P., Barros, N., & Carvalho, A. C. (2014). Can artificial neural networks be used to predict the origin of ozone episodes?. Science of the total environment, 488, 197-207.

[16]. Samadianfard, S., Delirhasannia, R., Kisi, O., & Agirre-Basurko, E. (2013). Comparative analysis of ozone level prediction models using gene expression programming and multiple linear regression. Geofizika, 30(1), 43-73.

[17]. Baawain, M. S., & Al-Serihi, A. S. (2014). Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network. Aerosol and air quality research, 14(1), 124-134.

[18]. Geng, F., Tie, X., Xu, J., Zhou, G., Peng, L., Gao, W., ... & Zhao, C. (2008). Characterizations of ozone, NOx, and VOCs measured in Shanghai, China. Atmospheric Environment, 42(29), 6873-6883.

[19]. Stathopoulou, E., Mihalakakou, G., Santamouris, M., & Bagiorgas, H. S. (2008). On the impact of temperature on tropospheric ozone concentration levels in urban environments. Journal of Earth System Science, 117(3), 227-236.

[20]. Sousa, S. I. V., Martins, F. G., Pereira, M. C., & Alvim-Ferraz, M. C. M. (2006). Prediction of ozone concentrations in Oporto city with statistical approaches. Chemosphere, 64(7), 1141-1149.

[21]. Bandyopadhyay, G., & Chattopadhyay, S. (2007). Single hidden layer artificial neural network models versus multiple linear regression model in forecasting the time series of total ozone. International Journal of Environmental Science & Technology, 4(1), 141-149.

[22]. Wang, W., Lu, W., Wang, X., & Leung, A. Y. (2003). Prediction of maximum daily ozone level using combined neural network and statistical characteristics. Environment International, 29(5), 555-562.

[23]. Nishanth, T., Kumar, M. S., & Valsaraj, K. T. (2012). Variations in surface ozone and NO x at Kannur: a tropical, coastal site in India. Journal of Atmospheric Chemistry, 69(2), 101-126.

[24]. Prybutok, V. R., Yi, J., & Mitchell, D. (2000). Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. European Journal of Operational Research, 122(1), 31-40.

[25]. Zhu, Y., Chen, C., Shi, J., & Shangguan, W. (2020). A novel simulation method for predicting ozone generation in the corona discharge region. Chemical Engineering Science, 227, 115910.

[26]. Arsić, M., Mihajlović, I., Nikolić, D., Živković, Ž., & Panić, M. (2020). Prediction of Ozone Concentration in Ambient Air Using Multilinear Regression and the Artificial Neural Networks Methods. Ozone: Science & Engineering, 42(1), 79-88.

[27]. Mo, Y., Li, Q., Karimian, H., Fang, S., Tang, B., Chen, G., & Sachdeva, S. (2020). A novel framework for daily forecasting of ozone mass concentrations based on cycle reservoir with regular jumps neural networks. Atmospheric Environment, 220, 117072.

[28]. AlOmar, M. K., Hameed, M. M., & AlSaadi, M. A. (2020). Multi hours ahead prediction of surface ozone gas concentration: Robust artificial intelligence approach. Atmospheric Pollution Research, 11(9), 1572-1587.

[29]. Ferreira, C. (2001). Gene expression programming: a new adaptive algorithm for solving problems. arXiv preprint cs/0102027.

[30]. Jodeiri Shokri, B., Dehghani, H., Shamsi, R. (2020). Predicting silver price by applying a coupled multiple linear regression (MLR) and imperialist competitive algorithm (ICA). 1(1):101-104.

[31]. Jodeiri Shokri, B., Dehghani, H., Shamsi, R. Doulati Ardejani, F. (2020). Prediction of acid mine drainage generation potential of a copper mine tailings using gene expression Programming-a case study. Journal of Mining and Environment, 11(4): 1127-1140.

[32]. Shakeri, J., Jodeiri Shokri, B., Dehghani, H. (2020). prediction of blast-induced ground vibration using gene expression programming (GEP), artificial neural networks (ANNs), and linear multivariate regression (LMR). Archives of Mining Sciences, 65 (2):317-335.

[33]. Dehghani, H. (2018). Forecasting copper price using gene expression programming. Journal of Mining and Environment, 9(2), 349-360.

[34]. Jodeiri Shokri, B., Ramazi, HR., Doulati Ardejani, F., Moradzadeh, A. (2014) A statistical model to relate pyrite oxidation and oxygen transport within a coal waste pile: case study, Alborz Sharghi, northeast of Iran. Environmental Earth Sciences, 71: 4693-4702.

[35]. Soleimani, M., Jodeiri Shokri, B. (2015). Defining chromite ore production trend by CCD method to reach sustainable development goals in mining sector, Iran. Mineral Economics, 28: 103-115.

[36]. Jodeiri Shokri, B., Ramazi, HR., Doulati Ardejani, F., Sadeghiamirshahidi, MH. (2014). Prediction of pyrite oxidation in a coal washing waste pile applying artificial neural networks (ANNs) and adaptive neuro-fuzzy inference systems (ANFIS). Mine Water and the Environment, 33: 146-156.

[37]. Doulati Ardejani, F., Rooki, R., Jodeiri Shokri, B., Eslam Kish, T., Aryafar, A., Tourani, P. (2013). Prediction of rare earth elements in neutral alkaline mine drainage from Razi coal mine, Golestan Province, northeast Iran, using general regression neural network. Journal of Environmental Engineering 139 (6), 896-907.