ISSN 1870-4069

Backpropagation Neural Network for the Prediction of PM_{10} Contamination Data

Daniel Cerna-Vázquez[1], Carlos Lino-Ramírez[1], Arnoldo Díaz-Ramírez[2], Francisco Mosiño[1], Miguel Ángel Casillas-Araiza[1], Rosario Baltazar-Flores[1], and Guillermo Mendez-Zamora[1]

[1]Tecnológico Nacional de México, Instituto Tecnológico de León, León, Gto., México
cerna.daniel.dc@gmail.com, carloslino@itleon.edu.mx,
jfmosino@gmail.com, miguel.casillas@gmail.com,
charobalmx@yahoo.com.mx, guillermomendez06@gmail.com
[2]Tecnológico Nacional de México, Instituto Tecnológico de Mexicali, Mexicali, B.C.
adiaz@itmexicali.edu.mx

Abstract. The prevention of respiratory diseases caused by high air pollution rates is an important issue in big cities, where industrialization and overpopulation cause an increase in allergenic particles that aggravate the disease of allergic rhinitis and asthma, especially in childhood. The problem lies in the disinformation of the population about air quality and the preventive measures to be taken in order to avoid deterioration in health. In this paper, data are monitored by a sensor network that registers the most abundant allergen, called PM_{10}, for the city of León, Guanajuato. An artificial neural network (ANN) with a supervised Backpropagation training is used to predict future data until a minimum error is reached. The proposed methodology generates efficient results, measured in the error of the solutions and in execution time.

Keywords: Sensor network, Artificial neural network, Backpropagation training, Climate data prediction.

1 Introduction

One of the main causes of respiratory diseases, which generates great concern for big cities, is the air emission of pollutants, caused by various human activities, including industry [1, 2]. The use of non-renewable resources in the production of energy, such as oil or coal, generates important emissions of sulfur dioxide (SO2), carbon monoxide (CO), among others. On the other hand, the means of transport used in daily life are another alarming source of pollution. A large part of these pollutants emitted into the environment is generated by cars [3]. According to the United Nations, there are currently around 7 billion people in the world [4, 5], which represents an enormous source of pollution, aggravating the problem as people tend to migrate to big cities or the cities continue to expand, which leads to a higher emission of pollutants that deteriorate air quality. According to data from the "Consejo Nacional de Población" (CONAPO), 72.3% of the population in Mexico lives in metropolitan areas. In addition, according to the United Nations (UN), in the next 10 years rural populations will begin to decline significantly [6]. All this causes that the health of the people, in the big cities, are going to deteriorate more and more. Allergens are those particles that can cause and/or aggravate allergies. The main allergens found in the air are ozone (O3), sulfur dioxide (SO2), carbon monoxide (CO), lead (Pb), particulate matter (PM2.5 and PM10), among others [7, 8].

Constant monitoring of air quality using sensor networks [9] allows people, who suffer allergies, to be informed about the environmental conditions, in order to take pertinent actions and avoid a deterioration in health. In the city of León, Guanajuato, there is the monitoring of air quality by the "Instituto de Ecología del Estado" (IEE). IEE has the "Sistema de Monitoreo de Calidad del Aire del Estado de Guanajuato" (SIMEG), a system that is made up by three fixed monitoring stations, distributed in the city of León. This paper makes use of the data generated by one of the three stations of the SIMEG, which is called "Cámara de la Industria del Calzado del Estado de Guanajuato" (CICEG) [10]. This station generates measurements of pollutant allergens PM, O3, NO2, SO2, NO2 and CO. These allergens (also Pb) fall into the category of major air pollutants [11].

In the present work, in order to obtain further information, as exact as possible, about the updated level of air quality in the environment, a Backpropagation Neural Network (BP-ANN) is used to predict the future data of allergens, obtained by the sensor network installed at the CICEG station. ANNs are a tool that has proved effective in predicting future data. In [12], a BP-ANN is used for the short-term prediction of wind energy. In [13], good predictions in the stock market are obtained.

The rest of this article is organized as follows: Section 2 presents theoretical concepts used in the elaboration of this work. Section 3 shows the methodology. Section 4 shows the results that were obtained and finally, section 5, shows the conclusions.

2 Theoretical Framework

2.1 Air Quality Index (AQI)

The AQI is an indicator of daily air quality, which shows how clean the air around us is, and what are the effects by such air quality. AQI measurements range goes from 0 to 500 ppb. These measurements are divided into five categories (Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy and Dangerous), that are determined depending on the level of pollutant in a higher proportion [14]. The AQI focuses on the measurement of five major air pollutants: SO2, CO, NOx, PMx and O3. The level of air quality, determined by the AQI [15], is generated by the Equation 1:

$$I_p = \frac{I_{HI} - I_{LO}}{BP_{HI} - BP_{LO}}(C_P - BP_{LO}) + I_{LO}. \quad (1)$$

Where I_p is the air quality index, C_P is the concentration of pollutant observed, I_{HI} is the breakpoint AQI greater than C_P observed, I_{lo} is the breaking point AQI less than C_P observed, BP_{HI} is the breakpoint of pollutant greater than C_P observed and BP_{LO} is the breakpoint of pollutant less than C_P observed. The AQI breakpoints are shown in Table 1, which are replaced in the Equation 1 to know the air quality index.

Table 1. AQI classification and breakpoints

Category	Minor Breakpoint	Major Breakpoint	Category color
Good	0	50	Green
Moderate	51	100	Yellow
Unhealthy for Sensitive Groups	101	150	Orange
Unhealthy	151	200	Red
Very Unhealthy	201	300	Purple
Dangerous	301	500	Brown

For the breakpoints of the pollutants, the air quality semaphore published in the “Informe de Estado y Tendencia de la Calidad del Aire Guanajuato 2014” was used [10]. The breakpoints for pollutants obtained by the CICEG station are shown in Table 2.

Table 2. Classification of the five main pollutants, sensed by the CICEG station, with its units of measures

Pollutant	PM_{10}	O_3	SO_2	NO_2	CO
Unit of Measurement	$\mu g/m^3$	Ppb.	Ppb.	Ppb.	Ppm.
Good	0-54	0-64	0-99	0-198	0-9
Satisfactory	55-74	65-69	100-109	190-209	9-10
Not Satisfactory	75-174	70-130	110-174	210-315	11-15
Bad	175-274	131-184	175-239	316-420	16-22
Very Bad	>275	>185	>240	>420	>22

2.2 Backpropagation Artificial Neural Network (BP-ANN)

Backpropagation was introduced in 1986, by Rumelhart, Hinton and Williams. This is a type of gradient descent [16], because it uses the calculation of the gradients of a neural network to adjust the weights [17]. Due to the advantages offered by this ANN modality, it is one of the most used [12].

The activation function used for the ANN is the sigmoid. This activation function has a good performance when the data for the training are positive, in a range of values between 0 and 1 [17]. The Equation 2 shows the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{2}$$

In order to comply with the network requirement, a normalization of the values was performed, placing them in a range between 0 and 1, through the Equation 3 [18]:

$$y_i = \frac{d - x_{min}}{x_{max} - x_{min}}. \tag{3}$$

Where d, is the data to normalize, x_{max} is the maximum value in the data series and x_{min} the minimum value. A total number of epochs was determined as the stopping criterion, in addition to an stop error [19], which is measure by the Equation 4:

$$Error = (y_i - f(x_i))^2. \tag{4}$$

Where y_i, are the known data, and $f(x_i)$ the data calculated by the BP-ANN.

2.3 Particulate Matter (PM_{10})

PM are inhalable particles, with diameters that are generally 10 micrometers and smaller. These particles can be breathed but remain on the nose, this can cause nasal obstruction among other inconveniences [20]. Figure 1 [21], shows an illustration of this pollutant.

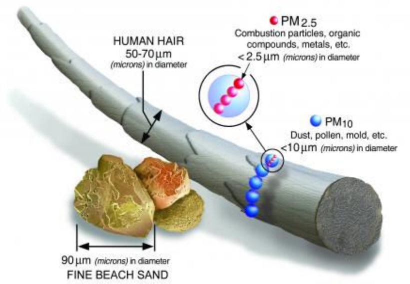

Fig. 1. Size of the PM_{10} particle

These particles are emitted directly from a source, such as construction sites, power plants, industries and automobiles. Also, are formed in the atmosphere as a result of complex reactions of chemicals. Exposure to such particles can affect both your lungs and your heart, and a variety of problems, as nonfatal heart attacks, irregular heartbeat, aggravated asthma, among others [21].

3 Methodology

This work was supported by the IEE, which facilitated the collection of the database generated by the CICEG station, located in the city of León. The database has 63, 913 records: from January 1, 2010 to March 31, 2017. In addition to the measurement of five pollutants (PM_{10}, O_3, NO_2, SO_2, NO_2 and CO), the database has the information of "Year, Month, Day, Time and Temperature". In this work, the data corresponding to the pollutant PM were taken, due to the fact that this pollutant is the most prevalent in the area, and dictates the AQI level for the CICEG station. Table 3 shows how the database is constituted.

Table 3. Database of levels for the five different contaminants, generated by the CICEG station

Date	Hour	O_3	SO_2	NO_2	CO	PM_{10}	Temperature
1/1/2010	0	7.02	10.40	151	3	362	14
1/1/2010	1	8.04	17.2	95	2.8	498	16
1/1/2010	2	10	18	101	3.3	490	17
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31/03/2017	21	9.24	16.1	90	2.14	82	22
31/03/2017	22	9.01	15.68	101	2.32	101.05	21.35
31/03/2017	23	8.77	15.06	85	1.86	90.34	20.98

First, the data from 8, 9 and 10 hours were selected, due to the pollutant was most present on that hours, and there is more urban mobility in the area. Figure 2, shows the behavior of PM10 pollutant measurements.

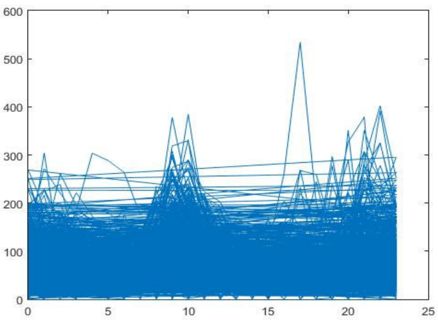

Fig. 2. Behavior PM_{10} pollutant measurements

We took one year of measurements for each hour; these correspond to data from March 1, 2016 to March 31, 2017. With these data we predict the level of pollutant of the selected hours to April 1, 2017. In total, we obtained 396 measurement data per hour. Figure 3, shows the methodology used.

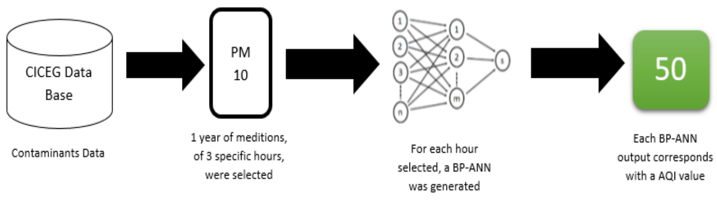

Fig. 3. Methodology used

For each hour selected for prediction, a BP-ANN was generated. Each of the values of the variables introduced in the input layer was normalized to satisfy the requirements of the network, through the Equation 3. The configuration that was used in the ANN is shown in the Table 4.

Table 4. Neural network configuration

Variable	Epochs	Entry/Hidden/Output Layer Neurons	Stop criterion	Learning rate	Weights
PM10	50000	316/310/1	0.001	0.03	-2 to 2

For training process, we took 396 data, which corresponds with 396 days of measurements from March 1, 2016 to March 31, 2017. Of these, 316 were taken for training; each value corresponds to an input neuron of the BP-ANN. The remaining 80 data were taken for testing, and thus performed the calibration of the weights. The output of the BP-ANN was compared with a value already known until reaching the stop criterion (0.001) or a specified number of epochs. We performed 35 runs, obtaining the median of the results, as a parameter to know the efficiency of the prediction.

This configuration shown in Table 4 was selected according to the best results given by previous experimentation, using different parameters. Once the value of the contaminant introduced into the ANN is obtained, the inverse process is done to the normalization, making use of the Equation 6, which is obtained by the Equation 3:

$$d = y * (x_{max} - x_{min}) + x_{min}. \tag{6}$$

Where y, is the normalized data, x_{max} is the maximum value in the standardized data set and x_{min} the minimum. Once the data is obtained in its original form, it is entered into the system through Equation 1, and thus knows the level of air quality. These levels of air quality are compared to each other to determine which the one that generates a higher AQI is.

4 Results

Once the network was trained, for the validation process, we add a vector whose output was not known by the network. The output not known, corresponded to the following day April 1, 2017 to those selected for training and testing. The validation data is known to verify the efficiency of the BP-ANN. Due the output of the network is a normalized data; the Equation 6 is used to know the real data and compare it. Table 5 shows the results of the experiments carried out in the prediction of the selected data from the CICEG station with the BP-ANN.

Table 5. Result of the 35 runs performed with the neural network

Hours	8	9	10
Real	53.71	88.96	83.19
Neural Network Output	52.80	87.45	85.10

The values of the pollutants were replaced in the Equation 1, in order to know which the one that determines the level of air quality is. Table 6 shows the results.

Table 6. AQI value for the results of the neural network experiments shown in Table 5

Hour	8	9	10
AQI	48.88	107.16	105.99
Category	Good	Sensitive Groups	Sensitive Groups

The behavior of the observed error in the training process of the BP-ANN, is observed in Figure 4.

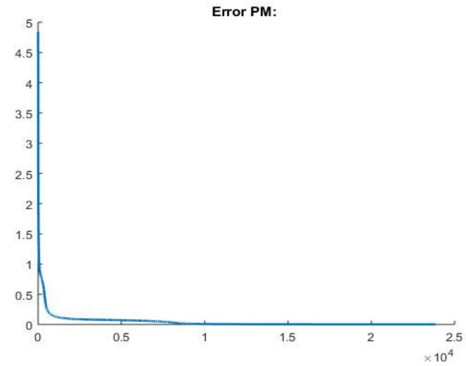

Fig. 4. Error behavior in the training process of the BP-ANN

As is shown in Table 6, the highest value AQI index was the PM_{10} for the 9 hours. Because the AQI index is taken on the basis of the pollutant that is most in the environment, the AQI predicted would be 107.16, which equates to the air quality category “Unhealthy for Sensitive Groups".

5 Conclusions

In this paper, the effectiveness of BP-ANN to predict has been demonstrated once again. As can be seen in Table 5, the values obtained at the output of the neural network are very close to the actual values (not known by the neural network) of the PM_{10} pollutant measured by the CICEG station. The system is effective, because the results allow having confidence in the predictions of the neural network.

It is important to mention that the values of the pollutant measurements obtained by the CICEG station vary greatly from one hour to another, due to factors such as traffic levels, work schedules in factories near the station, climatic conditions, among others. As this can vary (for better or for worse) from one hour to another, the system is also able to send alerts based on the observations of the station. It is expected with this to encourage people suffering from respiratory diseases, to be more aware of the air quality of the environment in which they develop. Based on the opinion of experts, the problem of pollution is something that will continue to exist and will continue to be a focus of attention for society, and by staying informed you can create awareness to take more care of the environment.

As future work, it is proposed to perform the processing of the databases generated by the other stations located in the city of León, and thus to have a general overview of the level of air quality in the city. Also, it is proposed to use other prediction tools such as least square data adjustment, and to perform experiments to see which offers better performance in prediction and computational cost or other features that help to streamline the prediction system. Besides, use sensors that allow us to know the quality of the air indoors, due that depending on indoor pollutants it can even lead to the generation of seasonal asthma and other diseases.

Acknowledgements. The financial support provided to the National Council of Science and Technology (CONACYT) is gratefully acknowledged for the accomplishment of this work. Also, special thanks to Dr. M. Ornelas-Rodríguez and engineers D. Duarte-Carrera and A. Godínez-Bautista, for their appreciated comments.

References

1. Yang, X., Du, J., Liu, S., Li, R., Liu, H.: Air pollution source estimation profiling via mobile sensor networks. In: 2016 Int. Conf. Comput. Inf. Telecommun. Syst., pp. 1–5 (2016)

2. Baralis, E., Cerquitelli, T., Chiusano, S., Garza, P., Kavoosifar, M.R.: Analyzing air pollution on the urban environment. In: IEEE 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1464–1469 (2016)
3. Bose, B.: Global Warming: Energy, Environmental Pollution, and the Impact of Power Electronics. IEEE Ind. Electron. Mag. 4, 6–17 (2010)
4. Guo, D., Zhang, Y., He, L., Zhai, K., Tan, H.: Chebyshev-polynomial neuronet, WASD algorithm and world population prediction from past 10000-year rough data. In: Proceddingd of the 27th Chinese Control Decis. Conf. CCDC 2015, pp. 1702–1707 (2015)
5. Melorose, J., Perroy, R., Careas, S.: World population prospects. United Nations. 1, 587–92 (2015)
6. Centro de noticias de la Organización de las Naciones Unidas. http://www.un.org/spanish/News/story.asp?NewsID=11754#.WNGX_Ts1-Um
7. Onatra, W., Vargas, S., Paez, E., Rojas, D., Lopez, A.: Correlation Between Acute Respiratory Disease (Are) in Pregnant Women and the Air Quality. Rev. U.D.C.A Act. Div. Cient. 12(2), 27–37 (2009)
8. Saad, S.M., Saad, A.R.M., Kamarudin, A.M.Y., Zakaria, A., Shakaff, A.Y.M.: Indoor air quality monitoring system using wireless sensor network (WSN) with web interface. In: 2013 Int. Conf. Electr. Electron. Syst. Eng., pp. 60–64 (2013)
9. Pieri, T., Michaelides, M.P.: Air pollution monitoring in lemesos using a wireless sensor network. In: IEEE 18th Mediterranean Electrotechnical Conference (MELECON). pp. 1–6 (2016)
10. Instituto de Ecología del Estado de Guanajuato. http://ecologia.guanajuato.gob.mx/sitio/calidad-del-aire/1/Sistema-de-Monitoreo-de-la-calidad-del-Aire-del-Estado-de-Guanajuato-(SIMEG)
11. Bagieński, Z.: Traffic air quality index. Sci. Total Environ. 505, 606–614 (2015)
12. Zhang, G.: Prediction of Short-Term Wind Power in Wind Power Plant based on BP-ANN. In: Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC*)*, pp. 75–79 (2016)
13. Guanqun, D., Fataliyev, K., Wang, L.: One-step and multi-step ahead stock prediction using backpropagation neural networks. In: 9th Int. Conf. Information, Commun. Signal Process, pp. 2–6 (2013)
14. Kyrkilis, G., Chaloulakou, A., Kassomenos, P.A.: Development of an aggregate Air Quality Index for an urban Mediterranean agglomeration : Relation to potential health effects. Sci. Direct. 33, 670–676 (2007)
15. García, B.: Los Índices de Calidad del Aire: Alcances y Limitaciones. Conciencia Tecnológica 42. 74–76 (2011)
16. Heaton, J.: Artificial Intelligence for Humans: Deep Learning and Neural Networks. Heaton Research Incorporated, 2015 (2015)
17. Heaton, J.: Neural networks editorial board addresses and specialities. Heaton Reasearch Incorporated, 2011 (2011)
18. Heaton, J.: Artificial Intelligence for Humans: Fundamental Algorithms. Heaton Research Incorporated (2013)
19. Kandel, A.: Introduction to pattern recognition: statistical, structural, neural, and fuzzy logic approaches. World Scientific (1999)
20. Cao, C., Jiang, W., Wang, B., Fang, J., Lang, J., Tian, G., Jiang, J., Zhu, T.F.: Inhalable Microorganisms in Beijing’ s PM 2.5 and PM 10 Pollutants during a Severe Smog Event. Enviromental Sci. Technol. 48, 1499–1507 (2014)

21. Environmental Protection Agency. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics