VOL. 11, NO. 11, JUNE 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

A GENERAL REGRESSION NEURAL NETWORK FOR MODELING THE BEHAVIOR OF PM10 CONCENTRATION LEVEL IN SANTA MARTA, COLOMBIA

R. Valencia, G. Sanchez and I. Diaz
Faculty of Engineering, Universidad del Magdalena, Santa Marta, Colombia
E-Mail: gsanchez@unimagdalena.edu.co

ABSTRACT

The presence of particulate matter in the air is a risk for human health, especially, when we are exposed to constant sources for a long time. In addition, this pollution factor causes infrastructure damages to the built environment, and damages to nature. Since these damages are permanent, some studies analyze the particulate material distribution over cities and estimate the Total Suspended Particle (TSP), which is a measure of the concentration level of particulate matter in a sampled area. The results are used to predict the behavior of particulate matter concentration, and, in this way, to implement contingency plan and prevent future effects. We present a methodology based on ANN (Artificial Neural Network) model to predict the particulate matter concentration. Initially, the method applies a stage of data preprocessing to smooth the time series, eliminates outliers, corrects missing data, and standardizes and normalizes the data. Then, the method trains a multi-layer perceptron model with a backpropagation algorithm. Finally, we estimate some error and accuracy measures to validate the predictions. The proposed ANN model obtained an accuracy between 85 to 95% in the experiments carried out in this work.

Keywords: artificial neural networks, PM10, air quality.

1. INTRODUCTION

High concentration of substances in the atmosphere as a consequence of human activities may have damaging effects on the environment and human health (Koening, 2000; Fukuda and Takaoka, 2007). These substances are formed of microscopic particles of solid or liquid suspended in gas. Carbon monoxide, sulfur dioxide, and volatile organic compounds are the main components of these pollutant substances that can be seen as dust, smoke, fog and ash in the atmosphere (Jiménez, 2001). The microscopic particles suspended in the atmosphere are also known as particulate matter and are commonly classified as coarse particle with a diameter of 10 micrometers (µm) or less, also called PM10, and fine particulate matter with a diameter of 2.5 µm or less also called PM2.5. According to the World Health Organization for each 10 $\mu g/m^3$ that the PM10 increases, the risk of death also increases in a 0.5%. Therefore, the presence of suspended particles in the air is considered an important risk to human health.

The impact of air pollution on human health is a topic that has received increasing attention from the governments and scientific community. The government has defined high standards and encouraged the acquisition of accurate and timely information to quantify the pollution level in the cities. Researchers on air quality process this information using statistical, geographical, and computational approaches (Fukuda and Takaoka, 2007). These computational approaches are mainly time-series analyzes and neural networks (NNs).

Wang *et al.* (2003) described a type of NN model based on Radial Basis Function (RBF) to predict the daily maximum ozone concentration level. The authors combined the NN model with the statistical characteristics of ozone and found a relationship between photochemical production and atmospheric accumulation of O_3. This model and the included characteristics yielded better predictions than the normal radial basis network. Previous work in this area is Gardner and Dorling (1999). Xu and Ho (2002) and Jiao *et al.* (2001) analyzed air pollution using a wavelet neural network which is a variant of the RBF network. According to the authors, this model presents some advantages in comparison to a general network model, such as faster convergence, adaptable structure, and avoidance of local minima.

Hajek and Olej (2012) introduced a prediction model for the daily average ozone level based on an NN, support vector regression, and uncertainty approach. The use of different methods made possible to compare prediction accuracy and stated some recommendations to micro-regional public administration management.

Recently, McCreddin *et al.* (2015) reported the result of an experimental assessment of personal exposure to PM10 in an office worker population. They compared the accuracy of different modeling techniques like ANN (Artificial Neural Network), Monte Carlo simulation, and time-activity based models, and concluded that the Monte Carlo technique produced the most accurate results according to three statistical measures of performance.

In contrast to a McCreddin *et al.* (2015), another work suggested that ANN provides better prediction accuracy than other typical techniques used for air pollution prediction (Vakili *et al.*, 2015). This last work predicted the daily absorption of global solar radiation on a land surface using an ANN model. The model used parameters like the daily maximum and minimum temperature, relative humidity, and wind speed. The contribution of Vakili *et al.* was the introduction of particulate matter into the model, yielding predictions that are more accurate.

Another approach to deal with time series prediction is the use of hybrid models. Feng *et al.* (2015) proposed a hybrid model that combines air mass trajectory and wavelet transformation. The purpose of this work was to improve the ANN accuracy to predict two days in advance the daily average concentrations of PM2.5. This work classified thirteen stations into dirty and clean air transportation using the air mass trajectory and the PM2.5 measures. This information was decomposed in sub-series with lower variability with wavelet transformation. According to the results, the hybrid model predicted PM2.5 concentration per days with an accuracy of 90% on average. Although NN models are prone to data over fitting, they have greater advantages for air pollution prediction compared with statistical methods. These advantages are due to air dynamics encompass multiple seasonality, long memory, and heteroscedasticity (Fasso and Negri, 2002; Zhang and San, 2004).

Our general purpose is to model the PM10 concentration level in Santa Marta, a city located in Colombia, a Latin-American country. Although this is a no industrialized city, it is an important port for coal exportation. The air quality in the city is deteriorated by coal contamination and traditional emission sources, such as motor vehicles, crops, factories, and civil works. The presence of particulate matter in Santa Marta gets worse due to the strong winds that displace the particles from the coastline to the urban area.

In compliance with national emission regulations stated by the Ministry of the Environment (Min Ambiente, 2010), local authorities monitor emissions across the city to ensure that the concentration levels of particulate matter do not exceed the limits causing damages to the environment and people's health. Garcia and Vergara (2014), who work in this field, determined the spatial and temporal variation of the STP and the breathable fraction (PM10) generated by the seaport activities. This work concluded that the air quality in Santa Marta is strongly affected by high levels of STP and PM10concentrations, especially in the north and south of the city. The authors also found some evidence indicating environmental problems that may cause public health issues in a long term.

Although Garcia and Vergara (2014) work provided valuable information to estimate pollution levels in the city, they revealed the damages, but not prevented them before happening. Thus, we are interested in implementing technological solutions to predict the pollution impact and allow the implementation of accurate contingency plan. In this context, works that have attempted to predict the air pollution concentration levels are Luna *et al.* (2014) and Wang *et al.* (2015).

2. METHODOLOGY

We implemented a methodology divided into two main phases. The first one includes data selection and time-scaled transformation, elimination of outliers using smooth transformation, and standardization and normalization of the data. The second phase consists of the ANN model design and parameter selection. Finally, we trained the ANN model to simulate the PM10 behavior (see Figure-1).

The data used for the time series were taken from twelve PM10 monitoring stations installed in the SVCA (*Sistema de Vigilancia de Calidad del Aire*) administered by the local entity Corpamag, in Santa Marta. The SVCA related information is a web public access data. Figure 2 shows the spatial distribution of the stations and their associated IDs.

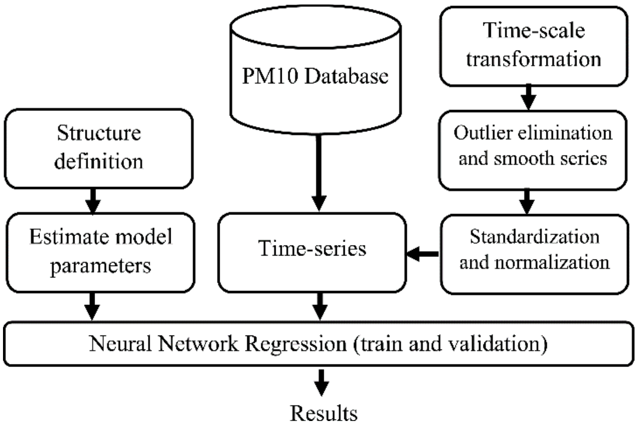

PM10 Database
Time-scale transformation
Structure definition
Outlier elimination and smooth series
Estimate model parameters
Time-series
Standardization and normalization
Neural Network Regression (train and validation)
Results

Figure-1. Block diagram summarizing the methodology.

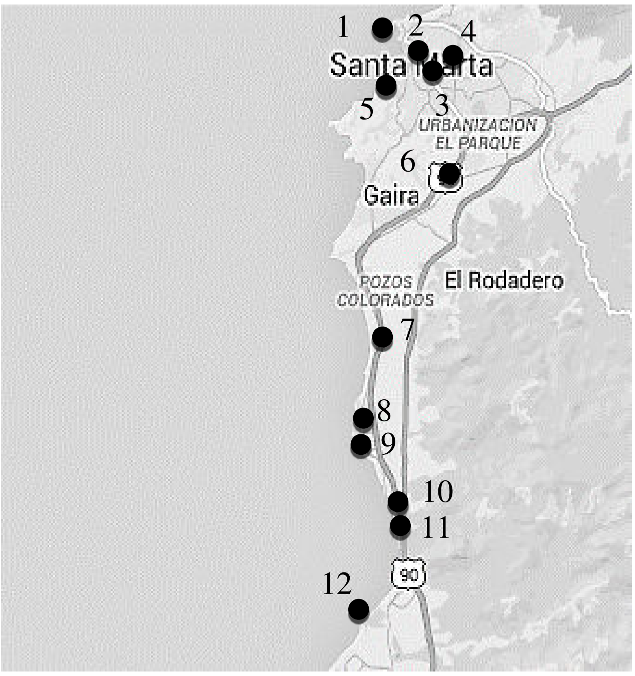


Figure-2. Air quality monitoring stations in Santa Marta city.

Time-scale transformation

The monitoring station measures the concentration of different particles like STP and PM10. Table-1 shows the variables measured by each station. We are concerned with the PM10 variable.

Table-1. Monitoring stations and measured variables.

Variable	Stations
PM-10	2, 6, 7, 9, 10, 11
STP	1, 3, 4, 5, 8, 10, 11, 12

The stations have data from the year 2007 to 2015. The data time-scale is three-day intervals. Because we wanted to model the monthly behavior, the time series were time-scaled to monthly values using the geometrical mean (Eq. 1). We selected the geometrical mean because itis little affected by the fluctuation of sampling and is robust to extreme values or outliers.

$$\bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i} \qquad (1)$$

where x_i is each of the samples of PM10 concentration measuredin a month, n is the total number of samples in a month, and $\bar{x}$ denotes the geometric meanof the samples.

Outlier elimination by a smoothing filter

We smoothed the time series by using a Savitsky-Golay filter, which is a convolution of the simplified least squares of the signal and is based on a polynomial regression of p degrees with at least $2n + 1$ equidistant points.Thus, the principle of $2n + 1$ sampled equidistant points (p_{-n}, ..., p_0, ..., p_n) represents samples of a p-degree polynomial additively with zero-mean measurement noise (see Figure-3). Mathematically, the Savitsky-Golay filter is defined as:

$$\hat{T} = \frac{1}{n}\sum_{i=-m}^{m} C_i T_{j+1} \qquad (2)$$

where $T$is the original value in the time series,$m$is the filter window size, n is the number of convoluting integers $n = 2m + 1$,$\hat{T}$ is the estimated value, and C_i is the coefficient for the i-th value of the filter.

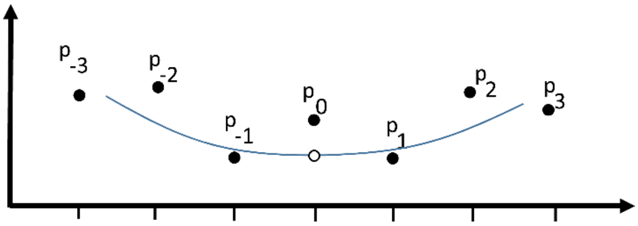


Figure-3. Savitzky-Golayfilter based on polynomial fitting.

Data standardization and normalization

After smoothing the data, the time-series is standardized and normalized. We standardize the time series by applying Equation 3:

$$T_{st} = \frac{\hat{T} - \mu}{\sigma} \qquad (3)$$

where T_{st} is the standardized data, $\hat{T}$ is the smoothed data (Equation 2), and $\mu$is the mean and $\sigma$is the variance of the time series. For normalization purposes, we applied equation 4:

$$T_{nor} = \frac{(T - min_T)(b - a)}{max_T - min_T} \qquad (4)$$

where $T_{nor}$is the normalized data, T the values to be normalized, $[min_T, max_T]$the range of the Tvalue, and$[a, b]$the range that will be reduced (in this case, [0,1]).

Definition of neural network structure

ANNs are computational models inspired by the biological neural networks of the brain. These models are composed of many nonlinear computational elements operating in parallel (Lippmann *et al.* 1987). A neural network learns a prediction model from a set of training data. Then, the model is able to respond to new patterns. The basic structure of ANNs consists of a set of inputs, a set of hidden units that conform hidden layers, a set of weights between hidden units, and an output. The inputs are processed in the hidden layers, where the weights play an important role to determine the output of each unit, until estimating the final output (Hecht, 1989). The main advantages of ANNs are that they can solve complex problems, generalize information, and yield accurate predictions.

During the training process, the set of weights in the ANN units is constantly updated in order to obtain an acceptable output, one that falls within an acceptable margin of error. The most common algorithms to carried out the training process are back propagation and forward propagation (Fei and Ling, 2007). Regardless of the training algorithm, the nature of the learning process is based on training patterns, which are simply input–output examples. The ANN error is the difference between the output predicted by the model and the actual output of each example. The number of neurons in each hidden layer depends on the underlying problem. Camargo (1990) stated that the more difficult the problem, the larger the sizes of hidden layers. Few neurons in the hidden layers may result in a high training error; while many neurons may result in overfitting to the training data. Other essential elements to design ANN models are the type of data, the learning function complexity, the activation function, and the training algorithm.

Usually, the NN structure is determined by trial and error or based on experience, and the number of neurons in the hidden layers is determined by some rule-of-thumb method that consists of applying the following equation:

$$N_{hn} = \frac{2}{3}(n + m) \qquad (5)$$

where $N_{hn}$is the number of neurons in the hidden layer, n the number of neurons in the input layer, and m the number of neurons in the output layer.

The backpropagation training algorithm updates the NN weights in two phases. In the first phase, the NN model receives the input data and processes them to generate an output. The algorithm calculates the error as the difference between the generated output and the desired data. In the second phase, the error is propagated in the opposite direction, from the output layer to the internal layers. Thus, each neuron receives only a portion of the error equivalent to a contribution percentage to the output. In this way, the neurons adjust their weights to reduce their error contributions (Zhang *et al.* 2015).

4. RESULTS AND DISCUSSIONS

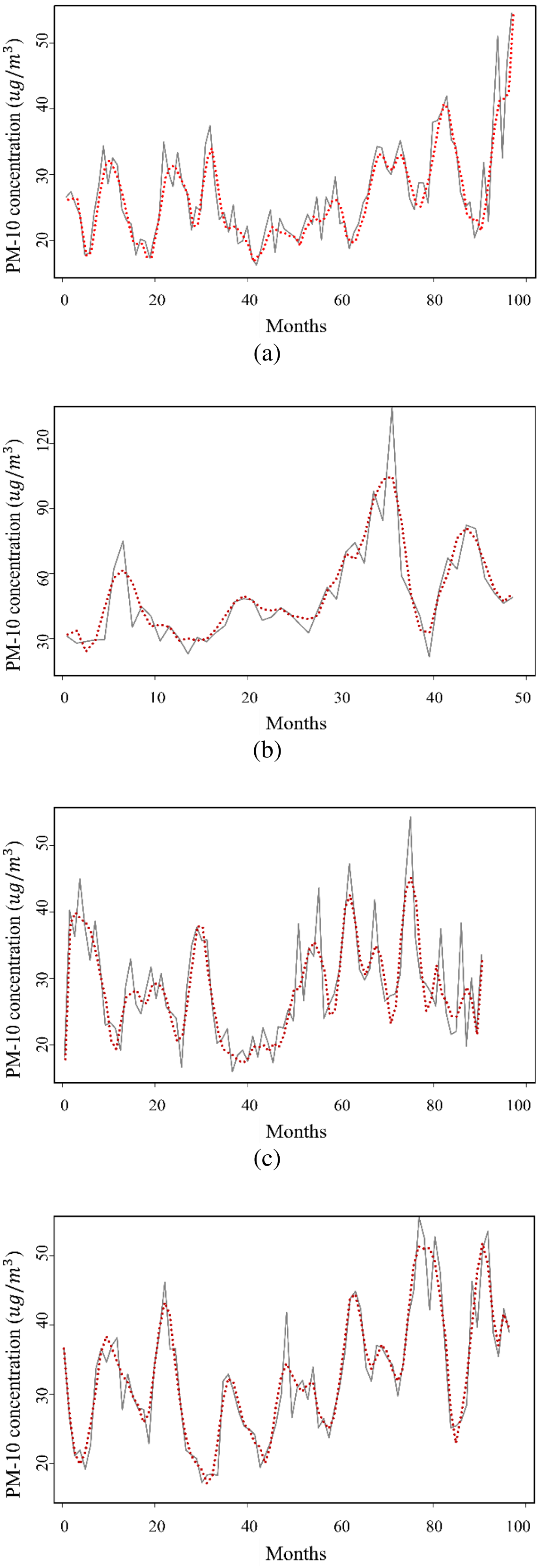
We evaluated the proposed methodology with data taken from the monitoring stations in the coastal area of Santa Marta city. We focused this work on the PM10 measures and used the geometrical mean to convert three-day measures to monthly measures. Due to the existence of flaws in the registration process, some station data were discarded for this work. We ended up using data from only four stations for these experiments.

First, we found the model parameters by carrying out a set of experimental tests. Then, we evaluated the accuracy of the ANN model to forecast a set of data. This kind of forecasting method is called the Univariate Forecast Model and is the model used for classical time series prediction (Box *et al.* 2008). The univariate forecasting method is based on the assumption that future values depend on some n set of preceding values. Thus,$x_{i+1} = f\left(\{x_j\}_n : j \leq i\right)$, where n is the number of preceding values, j is the index of values such as $j = i, i - 1, \ldots, i - (n - 1)$, and $f(\cdot)$ is the unknown function that fit over the preceding and future values.

Due to instrument malfunction or deficiencies in the data sensing procedure, the original set of data did not have the expected ten values measured monthly. Hence, we assumed that the measuring process of PM10 concentration generated data with some unknown noise level. For this reason, we applied a smoothing approach to the original time series. The smoothing approach was the Savitsky-Golay filter whose parameters were set to the polynomial degree $p = 7$ and $n = 13$.

Figure-4 shows some original time series with their respective smoothed versions. The result of the smoothing technique made it possible to reduce the effect of extreme values as well as abrupt changes that may be caused by the error model. Table-2 shows the statistics of the real-time series and smoothed time series. These values correspond to the monthly average of PM10 concentration level.

In the construction of the ANN model, the number of input parameters was established by experiments, and the number of hidden layers and neurons was estimated by$\frac{2}{3}(3 + 1)$. The model predicted future values based on three-days values, generating one output at once. Figure-5 shows the general structure of the ANN model used in this work.



(a)

(b)

(c)

(d)

Figure-4. PM10 concentration at selected monitoring stations. Continuous lines are original time series. The dotted lines are smoothed series.(a)–(d) correspond to station IDs 2, 6, 7, and10 respectively.

Table 2. Statistics of measured values and smoothed data (series), minimum and maximum values, mean and standard deviation (SD).

Station	Size	Series	Min.	Max.	Mean	SD
2	98	Real	17.36	55.73	28.14	7.27
		Smoothed	18.39	56.40	28.14	6.75
6	50	Real	20.77	121.85	45.22	19.20
		Smoothed	21.55	92.42	45.23	17.36
7	84	Real	14.53	51.94	27.20	7.75
		Smoothed	16.39	43.53	27.20	6.81
10	86	Real	14.82	62.3	33.71	11.30
		Smoothed	14.22	57.16	33.72	10.71

We used two different activation functions: a hyperbolic tangent function and a logistic function. Table-3 reports the steps needed for convergence as well as the error of the training process. The results of the ANN model over the data test of different time series, one for each station, are shown in Table-4. Although the logistic activation function needed significantly fewer steps to converge, the hyperbolic tangent function shows a minor error in all cases.

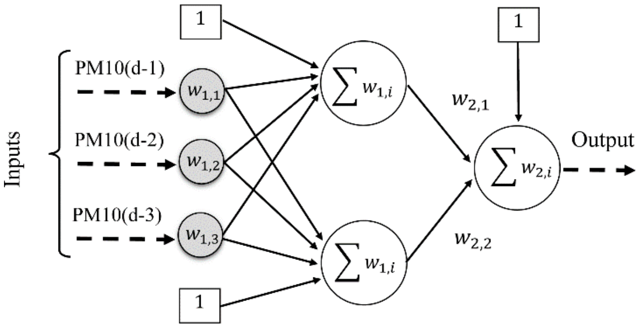


Figure-5. General artificial neural network architecture used for time-series regression.

Table-3. ANN results using different activation functions.

Station ID	Step size		Training error	
	Tanh	Logis	Tanh	Logis
2	153	228	0.060	0.075
6	3540	228	0.090	0.104
7	4848	439	0.158	0.200
10	1665	669	0.085	0.086

Error and accuracy measures

Toestimate the error rate, we computed two widely used error measures the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) (Feng *et al.*, 2015; Kun *et al.* 2008). Additionally, we computed an accuracy measure called the index of agreement (IA) during the experimental setup. IA was initially proposed as a statistical measure of the correlation between the real observed data and the predicted dataset. This correlation measure is an alternative to the correlation coefficient and coefficient of determination due to inadequacies in the use of such measures (Willmott, 1981).The equations of the computed measuresare defined by:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_{real} - T_{pred})^2} \tag{6}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|(T_{real} - T_{pred})\right|^2 \tag{7}$$

$$\text{IA} = 1 - \frac{\sum_{i=1}^{N}\left|(T_{real}-T_{pred})\right|^2}{\sum_{i=1}^{N}\left(\left|(T_{pred}-\bar{T}_{real})\right|+\left|(T_{real}-\bar{T}_{real})\right|\right)^2} \tag{8}$$

where N is the size of the test data, $T_{real}$the real time series, T_{pred} the set of outputs from the ANN model, and $\bar{T}_{real}$ the average of T_{real}.

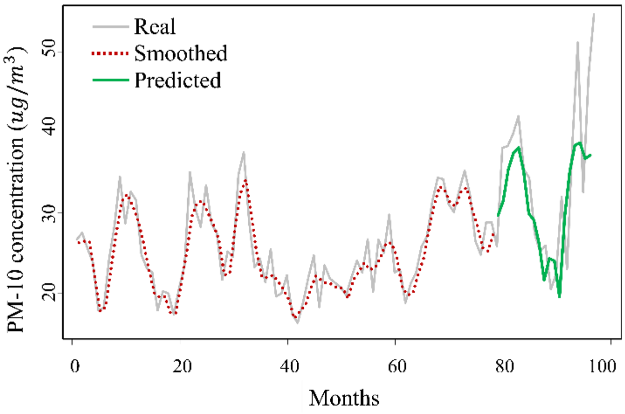
Table-4 reports the prediction errors for the validation data set. The hyperbolic tangent function produced a minor error for the four cases. The MAE measure yielded a similar behavior, reporting a minor error in all instances.

The analysis based on the index of agreement showed that the hyperbolic tangent function yielded better results than the logistic activation function. The hyperbolic tangent function generated values from this index from 81to 94%; while, the logistic function generated IA from 85 to 95%.

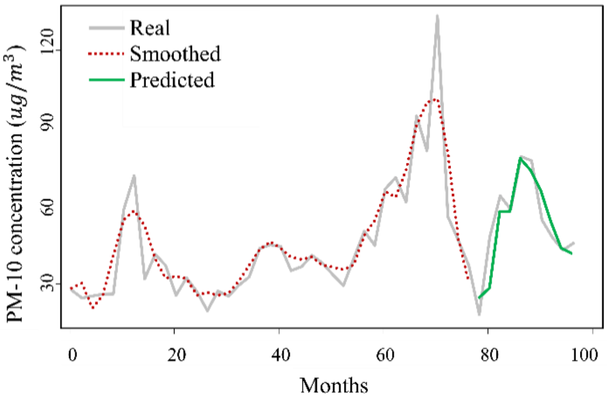
Figure-5 summarizes the ANN behavior over the validation data set. The green line shows the predicted values for each one of different time-series. Graphically, it displays the agreement between real and predicted values. The ANN model approximately the future behavior of the pollution behavior.

Table 4. Error and accuracy of the ANN model.

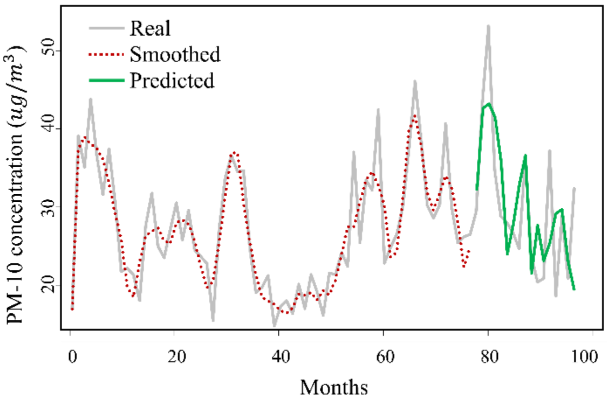
Station	RMSE		MAE		IA	
	Tanh	Logis	Tanh	Logis	Tanh	Logis
2	5.25	6.03	27.57	36.38	0.85	0.81
6	5.76	7.57	33.19	57.43	0.95	0.93
7	4.51	4.53	20.38	20.59	0.88	0.87
10	4.59	4.68	21.15	21.90	0.95	0.94



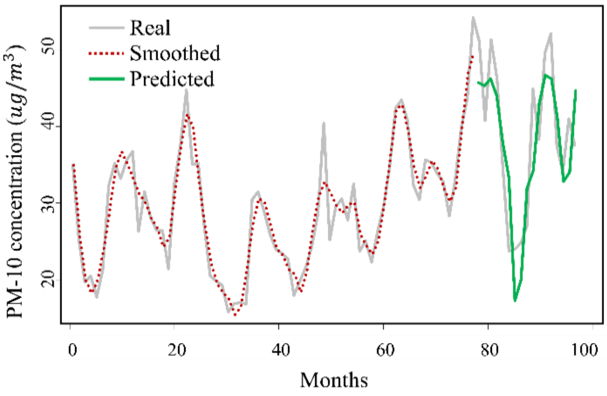
(a)



(b)



(c)



(d)

Figure-6. Real, smoothed training set, and the values predicted by the ANN model for each time series using tanh activation function. (a) - (d) correspond to station IDs 2, 6, 7, and 10, respectively.

4. CONCLUSIONS

We report aunivariate forecasting system for PM10 concentration level. The system is based onhistorical data and an ANN modelto predict the PM10 concentration levelin the following month. The data used in this work was taken from four monitoring stations located in Santa Marta city.

The ANN model designed for this work implemented two types of activation functions: hyperbolic tangent and logistic functions. The results showed that the ANN model had a better behavior with the hyperbolic tangent function. We analyzed the ANN's accuracy with different error measures and assumed some unknown noise level in the data obtained from the sensing station.

In the future, the introduction of an error model for the device used in the monitoring process will improve the representation of the data by the ANN model. Additionally, future studies should explore whether air pollution behavior can be most accurately modeled by including weather information. Although, weather information is not widely available. Future efforts should build a very robust monitoring infrastructure in Santa Marta city, to improve the actual information system.

REFERENCES

A.S. Luna, M.L.L. Paredes, G.C.G. de Oliveira, S.M. Corrêa. 2014. Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. Atmospheric Environment. 98, 98-104.

A. McCreddin, M.S. Alam and A. McNabola. 2015. Modeling personal exposure to particulate air pollution: An assessment of time-integrated activity modeling, Monte Carlo simulation and artificial neural network approaches. International Journal of Hygiene and Environmental Health. 218(1): 107-116.

Atakan Kurt, Betul Gulbagci, Ferhat Karaca and Omar Alagha. 2008. An online air pollution forecasting system using neural networks, Environment International. 34(5): 592-598.

Box G., Jenkins G. and Reinsel G. Time series analysis: forecasting and control. 4thed. Wiley, 2008.

Camargo F.A. 1990. Learning algorithms in neural networks. Working Paper, DCC Laboratory, Computer Science Department, Columbia University.

Fassò A. and Negri I. 2002. Non-linear statistical modeling of high-frequency ground ozone data. Environmentrics. 13(3): 225-241.

Fei Han and Qing-Hua Ling. 2007. A new learning algorithm for function approximation by incorporating a priori information into feedforward neural networks, in Third International Conference on Natural Computation2007 (ICNC 2007). 1: 29-33.

Gardner M.W. and S.R. Dorling. 1999. Neural network modeling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmospheric Environment. 33(5): 709-719.

Hájek P. and Olej V. 2012. Ozone prediction on the basis of neural networks, support vector regression and methods with uncertainty, Ecological Informatics. 12, 31-42.

Hecht-Nielsen R. 1989. Theory of the backpropagation neural network. Neural Networks, 1989. IJCNN. International Joint Conference on. Vol., No., pp. 593-605 vol. 1, 0-0.

Jiangshe Zhang, Nannan Ji, Junmin Liu, Jiyuan Pan, and Deyu Meng. 2015. Enhancing performance of the backpropagation algorithm via sparse response regularization, Neurocomputing. 153(4): 20-40, ISSN 0925-2312.

Jinhua Xu and Daniel W.C. Ho. 2002. A basis selection algorithm for wavelet neural networks. Neurocomputing. 48(1-4): 681-689.

Jiménez B. 2001. La contaminación ambiental en México: causas, efectos y tecnología apropiada. Colegio de Ingenieros Ambientales de México, A. C. México.

Koening J.Q. 2000. Health effects of ambient air pollution. How safe is the air we breathe? Boston: Kluwer Academic.

Kyoko Fukuda and Tadao Takaoka. 2007. Analysis of air pollution (PM10) and respiratory morbidity rate using the K-maximum sub-array (2-D) algorithm. In: Proceedings of the 2007 ACM Symposium on Applied Computing (SAC '07). ACM, New York, USA, pp. 153-157.

Licheng Jiao, Jin Pan and Yangwang Fang. 2001. Multiwavelet neural network and its approximation properties. IEEE Transaction Neural Networks. 12(5): 1060-1066.

Lippmann R.P. 1987. An introduction to computing with neural nets. in ASSP Magazine, IEEE. 4(2): 4-22.

Masoud Vakili, Saeed-Reza Sabbagh-Yazdi, Koosha Kalhor, Soheila Khosrojerdi. 2015. Using artificial neural networks for prediction of global solar radiation in Tehran considering particulate matter air pollution. Energy Procedia. 74, 1205-1212.

Ministerio de Ambiente, Vivienda y Desarrollo Territorial, MinAmbiente. 2010. RESOLUCIÓN 610 del 24 de marzo de 2010. Donde modifica la RESOLUCIÓN 601 DE 2006 abril 4, por la cual se establece la Norma de Calidad del Aire o Nivel de Inmisión, para todo el territorio nacional en condiciones de referencia. Bogotá, Colombia.

Organización Mundial de la Salud (OMS). Guías de calidad del aire de la OMS relativas al material particulado, el ozono, el dióxido de nitrógeno y el dióxido de azufre. Actualización mundial 2008. Washington: OMS.

Ping Wang, Yong Liu, Zuodong Qin, Guisheng Zhang. 2015. A novel hybrid forecasting model for PM10 and SO2 daily concentrations. Science of the Total Environment. 505, 1202-1212.

Saurabh Karsoliya. 202. Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. International Journal of Engineering Trends and Technology. 3(6): 714-717.

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

Willmott C. 1981. On the validation of models. Physical Geography. 2, 183-194.

Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, Jingjie Wang. 2015. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmospheric Environment. 107, 118-128.

Zhiguo Zhang and Ye San. 2004. Adaptive wavelet neural network for prediction of hourly NOX and NO_2 concentrations, In: Proceedings of Simulation Conference, Winter. 2: 1770-1778.