Research Article　　　　　　　　　　　　　　　　　　　　　　　　　　**Open Access**

Modeling of Surface and Weather Effects Ozone Concentration Using Neural Networks in West Center of Brazil

Amaury de Souza[1]*, Flavio Aristones[1] and Fabio Verissimo Goncalves[2]

[1]*Federal University of Mato Grosso do Sul, Institute of Physics, PO Box 549, CEP 79070-900, Campo Grande Mato Grosso do Sul, Brazil*
[2]*Federal University of Mato Grosso do Sul, Faculty of Engineering, Architecture and Geography, Graduate Program in Environmental Technologies, PO Box 549, CEP 79070-900, Campo Grande Mato Grosso do Sul, Brazil*

Abstract

The estimative of the concentration of surface ozone promotes the creation of data for planning forecasting the air quality, useful in the management of public health. The aim of this study was to develop an Artificial Neural Network (ANN) to estimate the concentration of surface ozone due to climate data daily. The ANN, the Feedforward Multilayer Perceptron kind, was trained taking as reference the daily concentration of ozone measured. In the intermediate and output layers we used activation functions type tan-sigmoid and linear, respectively. The performance of the ANN developed was very good, and it can be considered as part of the set of indirect methods to estimate the concentration of surface ozone. The proposed model can be used by the government as a tool to enable the public interventional actions during the period of atmospheric stagnation, when ozone levels in the atmosphere may represent risks to public health.

Keywords: Neural networks; Climate; Surface ozone; Public health; Air quality

Introduction

Surface ozone (O_3) is one of the most important pollutants in troposphere. Its concentration in any given area is the result of the combination of its formation, transport, destruction and deposition. The O_3 sources include: [1] photochemical reactions involving its precursors (volatile organic compounds and nitrogen oxides) with natural or anthropogenic origin; [2] downward transport from stratosphere; [3] long-range transport (intercontinental) of ozone from distant pollutant sources [1,2]. The increase of precursors' emissions due to the economic development of many countries in the world led to the rise of the surface O_3 concentrations [3-6]. Consequently, a public concern about its negative effects on human health, climate, vegetation and materials it has been observed [7-9].

About the human health protection, several studies were implemented to predict the O_3 concentrations [10-12]. The statistical models are the most commonly used ones, due to the complexity of the chemical chain reactions that are associated to O_3 formation and destruction. In this context, linear and nonlinear models have been applied to predict the concentration of this air pollutant. Multiple linear regression, principal component regression, quantile regression, among others, are a few examples of linear models [13-15] and on the other hand, artificial neural networks are the nonlinear models most commonly used [12,16-20]. Evolutionary procedures to determine predictive models were also applied, which include threshold autoregressive models optimized by genetic algorithms (GAs) and genetic programming models [21,22]. Moreover, in several research fields, GAs have been also applied to optimize data division, the weights or the structure of the artificial neural networks [23-26].

Data

Information on daily levels of ozone (O_3) were obtained from the Department of Physics of UFMS. The Ozone Analyzer which was used to perform the measurements has the working principle of the absorption of ultraviolet radiation by ozone molecule. The analyzer is installed near Campo Grande, away from local resources. The measurements are performed continuously 24 hours per day, and every 15 minutes, values are given of the ozone concentration. Then, when the arithmetic mean was calculated per day, it was assumed that this estimate was representative of air pollution in the city of Campo Grande. Information about rainfall, average temperature and relative humidity were obtained from Embrapa — Gado de Corte — Campo Grande.

In this study, we performed a descriptive analysis of variables which subsequently were associated with ozone concentration data, the rainfall climatic variables, maximum temperature, relative humidity and wind speed, from the period of 2004 to 2010.

Methods

Artificial neural networks

ANN can perform several functions such as classification, regression, association and mapping tasks [27-29]. They have a wide range of applications including adaptive control, optimization, medical diagnosis, decision making, as well as information, signal and speech processing [30]. ANN models are characterized by: (1) a set of processing neurons (also designated by nodes), (2) a pattern of connectivity among neurons, (3) an activation function for each neuron and (4) a learning rule. The processing neurons are distributed in layers: (1) input layer (first layer), (2) output layer (last layer) and (3) hidden layers (layers between the input and the output layers). The neurons in different layers are linked by synapses (each one storing a weight value) and the way which these linkages are done defines the

structure of the network. These models were described in more detail by [29,31].

In this study, a feed forward ANN with three layers was applied to predict surface ozone concentrations with five input variables (O_3, T, RH, speed, precipitation). A linear function was used as activation function of the output neuron. Concerning the hidden neurons, four functions were tested: sigmoid, hyperbolic tangent, inverse and radial basis. The early stopping method (training procedure is stopped when an increase of validation error is observed) was applied to try to avoid the over fitting.

Daily data were stored between January — 2004 and December — 2010, and the total were divided into a training group (2/3) and a test group (1/3). Ozone observed data were necessary for training and validation of the results.

The program for training and testing ANN was developed with Matlab software. Aiming the desired map, a lot of net topologies of the Feed Forward Multilayer Perceptron were tested with variations of the numbers of neurons of the intermediate layers. Since the air temperature, humidity, rainfall, wind velocity and the transport fleet are the main factors that influence the estimative of ozone concentration, its maximum, minimum and average values were used as input data in ANN. In the intermediate layer were used activation functions of tan-sigmoid type and in the output layer were used activation functions of linear type, featuring this neural net as a universal approximator of functions. The data standardization were made depending to the kind of activation function in the output layer of the RNA, this procedure became necessary. The software Matlab offers two forms of data standardization in an interval [-1,1] and with average=0 and variance=1 and finally the total data were divided in 2/3 consecutive for training and 1/3 for validation.

Considering that, in the beginning of training, the free parameters are randomly created and that these initial values could influence in the final result of the training, each net architecture was trained ten times, being selected that one presented the highest value of determination coefficient (R^2). This coefficient was calculated from the data of the observation of the ozone concentration in the test sample and the respective values estimated by ANN.

Aiming the desired map, were trained a lot of net topologies, varying the number of neurons, activation functions in the intermediate layers, as well the numbers of the interactions (Table 1).

The ozone values estimated by the ANN were compared with the numbers calculated by the accumulated percentage error, the Root Mean Square Error (RMSE), the exactitude coefficient of Willmot (d) and the performance index (c).

The RMSE was calculated from the equation1.

$$RMSE = \frac{1}{n}\sum_{i-1}^{n}\left(\tilde{Y}_i - Y_i\right)^2 \qquad (1)$$

According to Camargo and Sentelhas [32], the following statistics indicators are considered to correlate the values estimated with the measures: exactitude – index of Willmott "d"; and of trust or performance "c". The exactitude, related to the detachment of the estimated values in relation to the observed ones, is given statistically by the agreement index proposed by Willmott [31]. Its values varies from zero, for no one agreement, to 1, for the perfect agreement. The index is given by the equation 2:

$$d = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}\left(|P_i - O| - |O_i - P|\right)^2} \qquad (2)$$

Being: Pi = estimated value; Oi = observed value; O=average of the observed values.

The performance index "c", presented by Camargo and Sentelhas [32], evaluates the performance of the different methods of estimative. This index gathers the indexes of precision, given by the coefficient of correlation (r) that indicates the degree of dispersion of the obtained data in relation to the average, ie, the random error and of the agreement "d". The index "c" is calculated according the equation 3.

$$C = r.d \qquad (3)$$

Camargo and Sentelhas [32] proposed one criterion to interpret the performance of the estimative methods by the index "c", presented in the Table 2.

After the developing of the training algorithm of the ANN and the realization of analyses of the available climate data and the training algorithms, it was obtained an ANN capable of estimate, in a satisfactory mode, the concentration of surface ozone. This estimate is realized by mapping the relation between the maximum, average and minimum temperature data, maximum, average and minimum related humidity, wind speed, rainfall, the numbers of automotive vehicles that were counted as input and the concentration of reference ozone that is the desired output.

Results and Discussion

The ANN selected presented the best performance with the minimum configuration possible. This configuration is composed of one input layer with three variables, two intermediate layers each one with 4 and 2 artificial neurons, respectively, and one neuron in the output layer. The activation function of Sigmoid Hyperbolic Tangent type was adopted for the neurons in the intermediate layer. Generally, the trained nets presented better performances with smaller numbers of cycles with the ANN selected reaching better efficiency in 200 cycles. Beyond this it was verified that the nets with more than 200 cycles presented "memorization" problems.

The annual average value was c=0.81 with a great performance and an annual monthly average of performance equal to 0.79.

Parameter	Value
Number of neurons in the intermediate layers	1 to 5; 5 to 10
Activation functions in the intermediate layers	Sigmoid Logistic; Sigmoid Hyperbolic Tangent
Number of cycles	50; 100; 200; 500

Table 1: Parameters tested in the training of the RNAs.

C Value	Performance
>0.85	Great
0.76ᵃ 0.85	Very good
0.66ᵃ 0.75	Good
0.61ᵃ 0.65	Average
0.51ᵃ 0.60	Tolerable
0.41ᵃ 0.50	Bad
<0.40	Terrible

Table 2: Criterion of interpretation of the estimative performance of concentration of surface ozone.

Month	[O$_3$] obs	[O$_3$] est	RMSE	C
Jan	10.32	13.07	-0.27	0.88
Feb	12.08	13.55	-0.12	0.88
Mar	13.28	14.34	-0.08	0.84
Apr	13,88	14.08	-0.01	0.83
May	14.66	11.40	0.22	0.73
June	14.93	15.54	-0.04	0.69
July	15.54	20.50	-0.32	0.56
Aug	24.29	24.57	-0.01	0.87
Sep	29.69	26.30	0.11	0.86
Oct	26.79	21.19	0.21	0.79
Nov	21.20	19.84	0.06	0.82
Dec	16.10	14.49	0.10	0.68

Table 3: Statistics indicators of the adjust between the values observed of the ozone and the values estimated by then RNA, monthly average relative error, values of "c" from January of 2004 to December if 2010.

Table 3 presents the values of the performance index (c) and of the root mean square error (RMSE) to the ANN's. Lowest values of RMSE associated with highest values of "c" indicate the performance of the methodology in the estimate of ozone concentration from the collected data.

The ANN's developed generally presented a good performance, except in the month of July, when they presented statistics index RMSE of -0.32, presenting values of "c" with a terrible performance. The concentrations of ozone presented four months of great performance, as shown in Table 3.

The ANN's performance was very good, mainly due to lots of data used in its training, making its learning easier. It also contributed to the very good performance the fact that different architectures were tested in the network, i.e., different numbers of layers, algorithms of learning, number of cycles, etc.

Some works like [32,33], evaluated many architectures for the ANN's, obtaining exceptional performances. It was emphasized that the number of cycles used in the training of the ANN's was high, making its learning easier, reducing the possibility of memorization occurrence. The memorization leads ANN's to present a good statistic performance (a high value of "c" and low value of RMSE), because this one is calculated based only in the sample of available data. On the other hand, the memorization would lead to serious distortions in the spatialization of the concentration of ozone extremely high or extremely low.

Analyzing the values with the ANN's (Table 3) we can verify that the memorization didn´t occur, because were note evidenced severe deviations of the concentration of ozone estimated.

Analyzing the data of Table 2, we can verify that the average concentrations vary between 10.32 and 29.69 ppb (Table 3), with losing in the months of January, February and March, our rainfall season. The highest values were evidenced in the months of August, September and October, because it's the time to prepare the land for planting the crops.

We observed that high values of R² and "d" were obtained. This results were compared with those ones obtained in other studies with previsions of daily concentrations of ozone (Grivas, Chaloulakou [23] (0.60 and 0.86); Nagendra and Khare [34] (0.61 and 0.78)). The average of annual values of R² and "d" of this study were (0.8796 and 0.923798).

Figure 1 shows the graphic that compares the values observed and predicted by the model in the phase of validation. Figure 2 presents the histograms of the residues of the model evaluated in the phase

of validation. A good model must have a normal distribution of the residues, i.e., the histogram of the residue must be symmetric, in the shape of a bell. To visualize the performance of the model and of the ANN, the values observed and the simulations were compared as shown in the Figure 3. The graphical shows a good adjust of the model to the observed data, both in the phase of estimation/training and in the phase of validation.

Conclusion

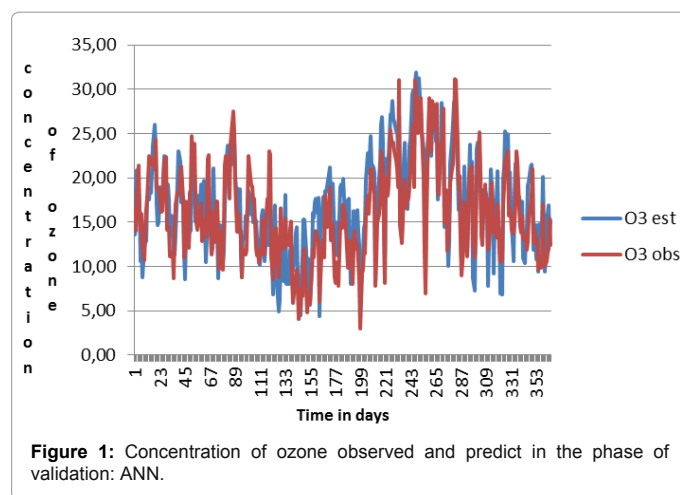The study of the methods for the estimate ozone concentrations

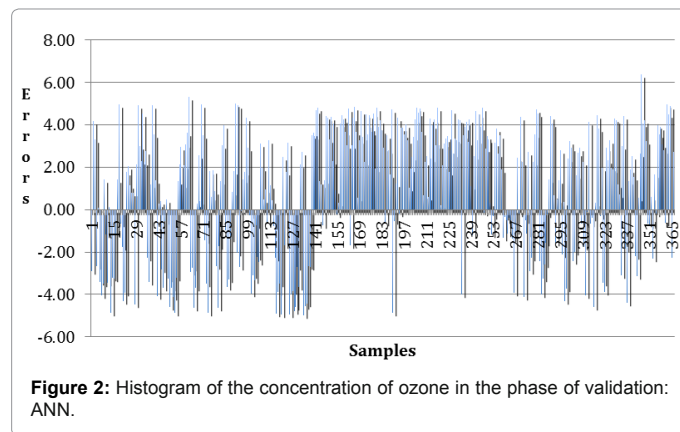**Figure 1:** Concentration of ozone observed and predict in the phase of validation: ANN.

Figure 2: Histogram of the concentration of ozone in the phase of validation: ANN.
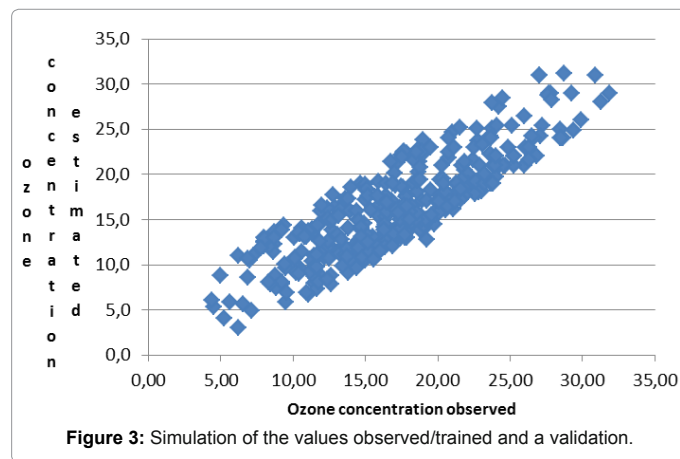

Figure 3: Simulation of the values observed/trained and a validation.

provides the average behavior of the parameters of study, which may be useful in prevention works of air quality, aiming the modeling work. Based on the results obtained in this work, we can conclude that:

1. The ANN's developed for estimating the ozone concentration presented a very good statistic performance.

2. There is a need of more training of the ANN's and variation of its architecture in order to obtain better statistic results.

3. The ANN's developed were capable of spatialize the concentrations of ozone without the presence of greater variances in its estimate.

4. Depending on the number of variables and the complexity of the architecture result the root mean square error may decrease or increase. Correlation values can be adjusted considering the size of the data presented to the network for training, variables that may represent more adequately modeled the environment as well as the development of other network architectures, enabling forecasts for longer periods.

Acknowledgement

References

1. Derwent RG, Kay PJA (1988) Factors Influencing the Ground Level Distribution of Ozone in Europe. Environ Pollut 55: 191-219.

2. EPA (Environmental Protection Agency) (1993) Air quality criteria for ozone and related photochemical oxidants. Environmental protection agency, pp: 3-06 (EPA-600/ P-93-004aF-cF).

3. Cartalis C, Varotsos C (1994) Surface ozone in Athens, Greece, at the beginning and at the end of the 20th century. Atmospheric Environment 28: 3-8.

4. Lisac I, Grubisic V (1991) An analysis of surface ozone data measured at the end of the 19th century in Zagreb, Yugoslavia. Atmospheric Environment 25: 481-486.

5. Nolle Michael, Ellul Raymond, Ventura Frank, Gusten Hans (2005) A study of historical surface ozone measurements (1884-1900) on the island of Gozo in the central Mediterranean. Atmospheric Environment 39: 5608-5618.

6. Roxanne Vingarzan (2004) A review of surface ozone background levels and trends. Atmospheric Environment 38: 3431-3442.

7. Jack Fishman (1991) The global consequences of increasing tropospheric ozone concentrations. Chemosphere 22: 685-695.

8. Fuhrer J, Skarby L, Ashmore MR (1997) Critical levels for ozone effects on vegetation in Europe. Environmental Pollution 97: 91-106.

9. Lippmann M (1991) Health effects of tropospheric ozone. Environmental Science & Technology 25: 1954–1962.

10. Hanna SR, Chang JC, Fernau ME (1998) Monte Carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. Atmospheric Environment 32: 3619-3628.

11. Robert Vautard, Matthias Beekmann, Jean Roux, Dominique Gombert (2001) Validation of a hybrid forecasting system for the ozone concentrations over the Paris area. Atmospheric Environment 35: 2449-2461.

12. Yi JS, Prybutok VR (1996) A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environ Pollut 92: 349-357.

13. JCM Pires, Martins FG (2011) Correction methods for statistical models in tropospheric ozone forecasting. Atmospheric Environment 45: 2413-2417.

14. Sousa SIV, Pires JCM, Pereira MC, Alvim-Ferraz MCM, Martins FG (2009) Potentialities of quantile regression to predict ozone concentrations. Environmetrics 20: 147-158.

15. Cannon AJ, Lord ER (2000) Forecasting summertime surface-level ozone concentrations in the Lower Fraser Valley of British Columbia: an ensemble neural network approach. Journal of the Air & Waste Management Association 50: 322-339.

16. Gardner M, Dorling S (2001) Artificial neural network-derived trends in daily maximum surface ozone concentrations. Journal of the Air & Waste Management Association 51: 1202-1210.

17. Inal F (2010) Artificial neural network prediction of tropospheric ozone concentrations in Istanbul, Turkey. Clean-Soil Air Water. 38: 897-908.

18. Latini G, Grifoni RC, Passerini G (2002) The importance of meteorology in determining surface ozone concentrations—a neural network approach. Ecology and the Environment 8: 405-414.

19. Lu HC, Hsieh JC, Chang TS (2006) Prediction of daily maximum ozone concentrations from meteorological conditions using a two stage neural network. Atmospheric Research 81: 124-139.

20. Jose CM Pires, Maria CM Alvim–Ferraz, Maria C Pereira, Fernando G Martins (2010) Atmospheric Pollution Research 1: 215-219.

21. JCM Pires, MCM Alvim-Ferraz, MC Pereira, FG Martins (2011a) Prediction of tropospheric ozone concentrations: application of a methodology based on the Darwin's theory of evolution. Expert Systems with Applications 38: 1903-1908.

22. Bowden GJ, Maier HR, Dandy GC (2002) Optimal division of data for neural network models in water resources applications. Water Resources Research 38: 2-1-2-11.

23. Chaloulakou A, Grivas G (2006) Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. Atmospheric Environment 40: 1216-1229.

24. Garcia-Gimeno RM, Hervas-Martinez C, de Siloniz MI (2002) Improving artificial neural networks with a pruning methodology and genetic algorithms for their application in microbial growth prediction in food. International Journal of Food Microbiology 72: 19-30.

25. Hansen JV, Mcdonald JB, Nelson RD (1999) Time series prediction with genetic algorithm designed neural networks: an empirical comparison with modern statistical models. Journal of Computational Intelligence and Electronic Systems 15: 171-184.

26. Corne SA (1996) Artificial neural networks for pattern recognition. Concepts in Magnetic Resonance 8: 303-324.

27. Mukta Paliwal, Usha A Kumar (2009) Neural networks and statistical techniques: a review of applications. Expert Systems with Applications 36: 2-17.

28. Schneider Gisbert, Paul Wrede (1998) Artificial neural networks for computer-based molecular design. Progress in biophysics and molecular biology 70: 175-222.

29. Gupta RR, Achenie Lek (2007) A network model for gene regulation. Computers and Chemical Engineering 31: 950-961.

30. Guoqiang Zhang, B Eddy Patuwo, Michael Y HU (1998) Forecasting with artificial neural networks: the state of the art. International Journal of Forecasting 14: 35-62.

31. Cortj Willmott et al. (1985) Statistics for the evaluation and comparison of models. Journal of Geophysical Research Atmospheres 90: 8995-9005.

32. Camargo AP, Sentelhas PC (1997) Evaluation of the performance of different potential evapotranspiration estimation methods in the state of Sao Paulo, Brazil. Journal of Agrometeorology 5: 89-97.

33. Moreira MC, Cecilio RA, Silva KR (2007) Comparison of methods for estimating the air temperatures in the Brazilian Northeast. In: Brazilian Congress Agrometeorology, 15,Aracaju. Anais. Aracaju : SBAgro, (CD-ROM).

34. Nagendra SMS, Khare M (2005) Modelling urban air quality using artificial neural network. Clean Technologies and Environmental Policy 7: 116-126.