Prediction of TSP concentration in a metallurgical city of Brazil using neural networks

M. M. C. Lima
Research and Development Centre, Usiminas, Brazil

Abstract

The aim of this study was to predict Total Suspended Particulate concentration (TSP) in the main areas of Ipatinga, a metallurgical city located in Minas Gerais state, southeast of Brazil. Artificial neural networks (ANN) were the modelling tool used. This model is able to predict pollutant concentration just by training the input and output parameters. The input parameters were meteorological such as wind direction, wind speed, rain, and ambient temperature and also seasonal such as, summer and winter. The output parameter used was the historical data of the total suspended particulate concentration taken between 1996 and 2004. In the modelling, the multilayer perceptron (MLP) model was tested. Among the MLP configurations evaluated, the topology 13-7-6 was chosen. The validation of the model was done by comparing the simulated with the observed values. The results of this model were also compared with the industrial source complex short-term dispersion model (ISCST3). The four statistical tools used to evaluate the fitting were mean squared error (MSE), fractional bias (FB), index of agreement (IA) and linear correlation coefficient (R). Comparing the results it was seen that the predicted values were better in some boroughs and were overestimated in others. Besides, the predicted results of the ANN model were better than the ISCST3 dispersion model.

Keywords: artificial neural networks modelling, multilayer perceptron, total suspended particulate concentration, prediction, ISCST3 dispersion.

1 Introduction

This paper introduces the study of prediction of Total Suspended Particulate concentration (TSP) in Ipatinga city using artificial neural networks.

WIT Transactions on Ecology and the Environment, Vol 147, © 2011 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR110101

Ipatinga is located in the Vale do Aço Region, in Minas Gerais state of Brazil where the main industrial activity is iron and steel making. In the iron and steel making process, the raw material handling and fuel combustion are the main causes of particulate emission. Depending on the particulate concentration, the air quality can be modified and causes a lot of damage, such as increasing the dust in residential areas, visibility impairment and harmful to health effects (USEPA [1]).

Mathematical models are often used to estimate environmental impacts, saving money from air quality monitoring. One of the most important features of models is the ability to predict or simulate future impact scenarios.

The dispersion or diffusion models have been traditionally applied to atmospheric mathematical modelling. These models are able to predict the pollutant concentration using mass balance of statistical data (pollutant emission, wind direction and speed, ambient temperature) and introduce a Gaussian mathematical equation as a solution of pollutant dispersion. Mitkiewicz [2] predicted TSP concentration in Ipatinga using industrial source complex short-term dispersion model (ISCST3).

Recently, artificial neural networks model (ANN) has been used in modelling complex problems. ANN, such as the ISCST3 model is able to predict air pollutant concentration just by training a set of input and output variables. It offers a mathematical solution by adjusting the weights in such a way that output will be close to real data. Comparing ANN to ISCST3 models, the first one has the advantage in adapting few variables to evaluate air pollutant dispersion. It can be seen in many papers published about this subject: Linyan and Wang [3], Wal and Janssen [4], Perez and Reyes [5], Viotti et al. [6], Zickus et al. [7], Perez and Reyes [8], Podnar et al. [9], Ordieres et al. [10], Hooybergh s et al. [11].

There are a variety of artificial neural networks models being used in modelling. Among them, the multilayer perceptron (MLP) is the most cited in air dispersion modelling (Gardner and Dorling [12]). The MLP structure consists of processing elements and connections. The processing elements, called neurons, are arranged in layers such as an input layer, one or more occult layers and an output layer (Haykin [13]). They are all interconnected.

In this context, this study aimed to develop ANNs using MLP. The input parameters were meteorological data and the output was the TSP concentration data. This model was also compared with the ISCST3 model. Specifically, it intended to: a) predict TSP concentration using ANNs in six air quality monitoring stations distributed in Ipatinga, b) determine the main variables responsible for the measured TSP concentration; c) investigate the behaviour of the created ANN due to different configuration proposals; d) validate the ANN model using the comparison between the simulated and measured values and e) compare simulated results between the ANN and ISCST3 models.

2 Material and methodology

2.1 Sampling

TSP concentrations were collected weekly between 1996 and 2004 from six air quality monitoring stations using High Volume sampler (HiVol) distributed in six districts around Ipatinga named: CA, BR, BA, NC, EC and CS. Meteorological variables (wind speed (m/s), wind direction, rain volume (mm), ambient temperature (^{o}C)) were collected hourly during the same period and converted to daily variables. The wind directions evaluated were north (N), south (S), east (E), west (W), northeast (NE), southeast (SE), southwest (SW) and northwest (NW). Further variables were also created as calm hours (wind speed less than 1 m/s) and seasonal cycle (winter and summer) during the same period from the same database to evaluate their effect on the TSP concentrations. The database contained 400 data.

2.2 Artificial neural modelling

Mathematical routine was developed using the software Matlab [14]. MLP was the artificial neural network model used. It had been tested lots of models with different configurations. The best topology was defined by the MSE (mean square error) analysis. This statistical analysis is suggested by Zhang et al [15] as being efficient in evaluating the artificial neural networks models. The meteorological and seasonal variables were used in the input layer. The choice for the model type and the input variables were due to a lot of published studies in literature as mentioned above. Two models with thirteen and six input data were evaluated. The thirteen data were: daily eight wind directions frequency, daily mean wind speed, daily mean rain volume, daily mean temperature, daily calm hours and seasonal cycle (winter = 1 and summer = 2) frequencies. The six were obtained using the principal component analysis (PCA). The purpose of this analysis is to obtain a small number of linear combinations which account for most of the variability in the data (Haykin [13]). In this case, six components had been extracted from the thirteen input data and they were evaluated in the modelling as input data. TSP concentrations, measured in the six monitoring sites, were introduced in the output layer. Only one occult layer was considered in the modelling. The number of neurons in the occult layer was changed as suggested by Zhang et al [15] and Kóvacs [16]. First, an exploratory data analysis was made, detecting if there were outliers in the database. After that the data were normalized and separated in “training” and “validation” sets. The training set was equivalent to 80% of the database and the validation set to 20% (Zhang et al [15]). The learning algorithms used during the training test were Levenberg-Marquardt and Backpropagation. The early stopping criterion was applied to stop the training. In the mathematical routine, expected error, initial weight, bias, activation function, learning algorithm, momentum term and the iteration cycles were established as usual. Finally, after comparing the real values to the simulated ones by MSE analysis, the best topology was defined.

2.3 Performance evaluation

The model validation was done by comparing between the real and simulated values. These simulated values were the model results obtained from the validation set. The average percentage error (E) was determined as shown in eqn. (1).

$$E = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|C_r - C_e|}{|C_r|}\right) * 100 \tag{1}$$

where n is the sampling size, C_e, C_r simulated and real values, respectively.

The tendency of the simulated results was evaluated by quantile-quantile plot and type I (false negative) and II (false positive) error analyses. In the type I error, the model under predicts the values, when they were supposed to be above a critical value. In this case, the critical value (TSP concentration) was the air quality standard of 80μg/m^3 applied in Minas Gerais state. In the other hand, in the type II error, the model over predicts the values. So they were supposed to be below a critical value.

The cluster analysis was applied to verify the effect of input data in the output data. This multivariate statistical technique aims to classify the observations or variables due to their similarity, applying a distance measure algorithm.

2.4 Comparison between ANN and ISCST3 models

The comparison between ANN and ISCST3 models was made. For running ISCST3 model, it was collected the meteorological parameters, topography and air emission sources characterization. The comparison among the two simulated results and real values, registered in the six monitoring sites, was made. The statistical evaluation tools used were linear correlation coefficient (R), mean square error (MSE), mean fractional bias (FB) and mean index of agreement (IA) (Olesen [17]).

The linear correlation coefficient is shown in eqn. (2):

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(C_e - \overline{Ce}\right)\left(C_r - \overline{Cr}\right)}{\sigma_{C_e}\sigma_{C_r}} \tag{2}$$

The mean square error is shown in eqn. (3):

$$MSE = \frac{\sum_{i=1}^{n}\left(C_r - C_e\right)^2}{n} \tag{3}$$

The mean fractional bias is shown in eqn. (4):

$$FB = \frac{1}{n}\sum_{i=1}^{n}\frac{C_e - C_r}{0.5\left(C_e + C_r\right)} \tag{4}$$

The mean index of agreement of (IA) is shown in eqn. (5):

$$IA = \frac{1}{n}\sum_{i=1}^{n} 1 - \frac{\left(C_e - C_r\right)^2}{\left[\left|\left(C_e - \overline{C_r}\right)\right| + \left|\left(C_r - \overline{C_r}\right)\right|\right]^2} \tag{5}$$

where $\overline{Ce}$: averaged simulated concentration, $\overline{Cr}$: averaged real concentration, n is the sampling size, C_e, C_r simulated and real values, respectively, σ_{C_r} : real concentration standard deviation, σ_{C_e} : simulated concentration standard deviation.

3 Results

The best ANN configurations results considering the MSE analysis are shown in table 1.

Table 1: MLP models results.

Model	Neurons			Alg.	MSE	Output neurons					
	Input	Occult	Output			CA	BA	BR	EC	NC	CS
1	13	7	6	LM	466.7	+	+	+	+	+	+
2	13	27	6	BP	636.5	+	+	+	+	+	+
3	13	14	5	LM	354.7	+	+	+	+	-	+
4	13	19	1	LM	251.5	+	-	-	-	-	-
5	13	10	1	LM	385.1	-	+	-	-	-	-
6	13	7	1	LM	208.4	-	-	+	-	-	-
7	13	7	1	LM	205.7	-	-	-	+	-	-
8	13	10	1	LM	911.3	-	-	-	-	+	-
9	13	9	1	LM	336.2	-	-	-	-	-	+
10	6	4	6	LM	509.1	+	+	+	+	+	+
11	6	13	1	LM	323.5	+	-	-	-	-	-
12	5	8	1	LM	386.4	-	+	-	-	-	-
13	6	6	1	LM	220.9	-	-	+	-	-	-
14	6	10	1	LM	188.6	-	-	-	+	-	-
15	6	4	1	LM	869.5	-	-	-	-	+	-
16	6	3	1	LM	382.0	-	-	-	-	-	+

(Alg.) Algorithm, (LM) Levenberg Marquardt, (BP) Backpropagation, (+) simulated, (-) not simulated.

Few ANN models had in the occult layer a half of neurons of the input layer while other ones (models 2 and 11) had more than double of neurons of the input layer as commented by Kóvacs [16]. The results obtained from models 1, 2 and 10, considering 6 neurons in the output layer, had the same order of magnitude. Leaving the NC air monitoring site out of evaluation (models 1 and 3), the MSE was reduced. Models 4 to 9 and 11 to 16 were created to evaluate each result from each air monitoring station. The results were very similar except for NC.

Model 12 had five input data due to PCA results. The results obtained from models 8 and 15 were probably due to other variables that were not introduced in the model, since the used variables were not able to explain entirely the TSP concentrations. Using MSE approach for comparison results between model 1 and 2, it was showed that Levenberg-Marquardt algorithm was better than Backpropagation. On balance, the MSE results among the models were very similar with the same order of magnitude. For this reason model 1 was chosen to describe the subsequent results. Comparing between models 1 and 10, the use of principal components as input data did not altered the MSE results significantly.

3.1 Performance evaluation

The average percentage errors for CA, BA, BR, EC, NC, and CS air quality monitoring sites were 35%, 24%, 27%, 31%, 42% and 37%, respectively. BA simulated results showed a good agreement with measured values. NC had the worst results. In regular meteorological conditions, BA is used in suffering particulate emissions from Usiminas more than the other sites, as it is located downwind from it. So the simulated results were better in BA than in the other ones. The result obtained in NC was the worst due to other variables that were not introduced in the model. Mitckiewicz [1] had also got the worst simulated results in NC using ISCST3 modelling.

The cluster analysis is shown in fig. 1. Analysing the distance measurement it is possible to identify three groups.

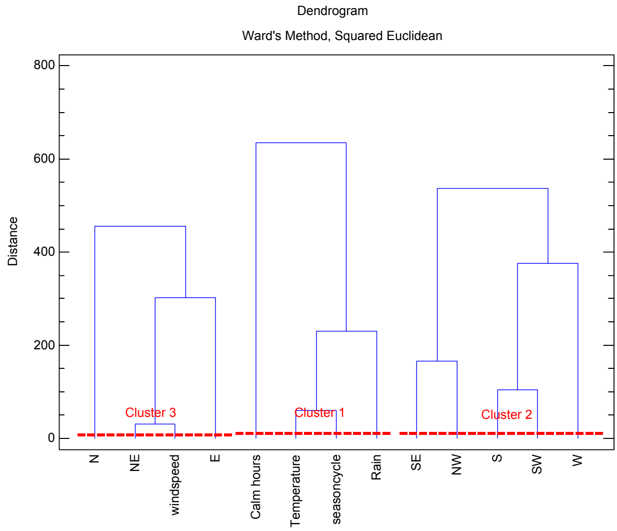

Figure 1: Cluster analysis.

Cluster 1, representing 50% of the output data, was characterized by N, NE, E wind directions, wind speed less than 1m/s, strong raining storms, ambient temperature about 21^{o}C and the presence of two seasonal cycles. In other words, 50% of TSP concentrations results were grouped due to similar input data mentioned above. Cluster 2 grouped 19% of the output data due to input variables: SE, S, NW, SW, W wind directions, wind speed more than 1,1m/s, weak raining, ambient temperature about 21°C and winter cycle. Finally, cluster 3 (24% of output data) was characterized by N, NE, and E wind directions, wind speed more than 2,1m/s, frequent raining, ambient temperature about 25°C and the presence of summer cycle. The analysis of quantile-quantile plot and type I and II error was made for all air quality monitoring sites, but in this paper, it will be only shown BA and NC results (figs. 2 and 3). The other ones can be seen in more details in Lima, M. M. C. [18].

In fig. 2, according to quantile-quantile plot analysis, the tendency of 52% of the predicted values was to overestimate the real values. They occurred in meteorological conditions from cluster 1. Considering the analysis of false positive and false negative errors, both were verified. But, the type I error was more often and was determined by the variables characterized by cluster 3. One hypothesis that could probably explain is the location of main particulate emission sources from Usiminas in relation to BA air quality monitoring site. Under those meteorological conditions (characterized by cluster 3), BA air monitoring quality site was downwind from them and if there was an emission increase during that period, the model was not able to evaluate it, as they were not introduced in the modelling. It could explain why predicted values was lower than real.

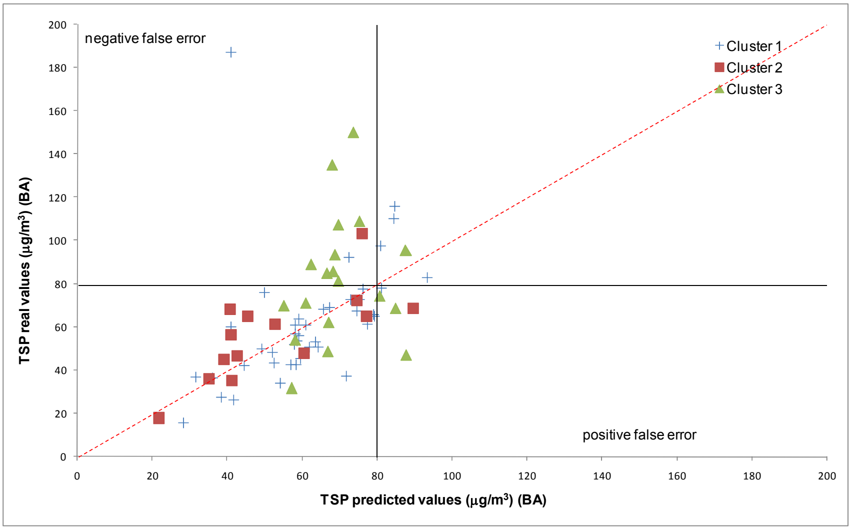

Figure 2: Comparison between real and simulated values for BA.

According to fig. 3, quantile-quantile plot analysis showed that a great part of simulated values was overestimated if compared to the real values. They also occurred in meteorological conditions described in cluster 1. The type II error was more common than the other one.

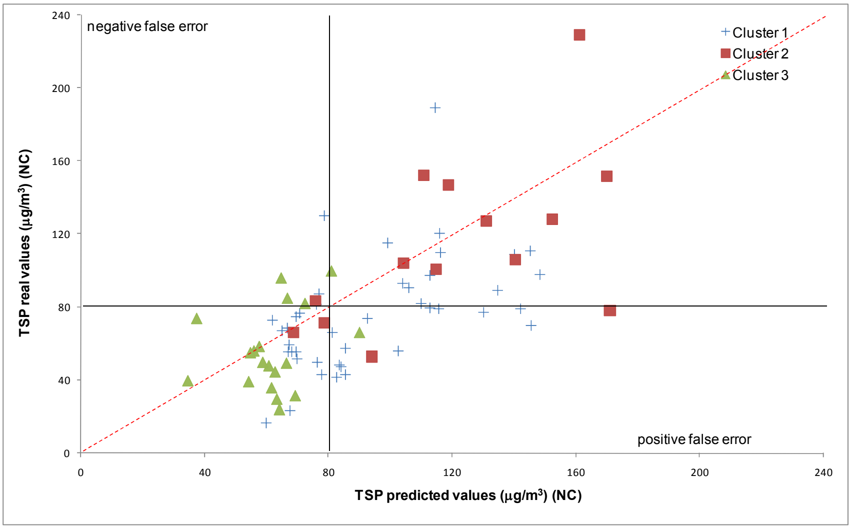

Figure 3: Comparison between real and simulated values for NC.

The possible cause was its location is not favourable and exposed to extra contributions. NC site is close to a paved road with heavy traffic. This situation was not modelled.

3.2 Comparison between ANN and ISCST3 model

The comparison between ANN and ISCST3 model results is shown is table 2.

Table 2: Statistical analysis results.

Site	Statistical analysis							
	FB		IA		R		MSE	
	ISCST3	ANN	ISCST3	ANN	ISCST3	ANN	ISCST3	ANN
CA	-0.29	0.06	0.36	0.66	0.02	0.48	1754.86	334.79
BA	-0.06	-0.07	0.53	0.62	0.37	0.47	2944.57	671.00
BR	0.42	0.12	0.21	0.69	0.02	0.56	6124.87	173.68
EC	-0.61	0.11	0.19	0.70	0.18	0.54	1652.21	202.97
NC	-0.82	0.15	0.53	0.76	0.44	0.65	4003.12	1027.18
CS	0.29	0.04	0.27	0.61	0.08	0.44	14616.33	390.40

FB and MSE usually measure bias of a model. R is the correlation between the observed and simulated values and IA shows the degree that the model predictions are error free. The ideal model would result in values of MSE = 0, R = 1, FB = 0 and IA = 1. Analyzing the mean fractional bias results, NC, BR and EC air quality monitoring stations had the worst results in both models. The mean index of agreement (IA) showed the EC and CS results were the worst in ISCST3 and ANN models, respectively. The values of IA and FB obtained in

ANN were better than the ISCST3 dispersion modelling. The values of R and MSE obtained in ANN were also better than the ISCST3 model. For this reason, ANN model should be considered closer to the ideal model.

4 Conclusions

This study showed that ANN can be a powerful data analysis tool to evaluate air pollutant dispersion. Even though, in the modelling, particulate emissions from Usiminas were not introduced as input variable, ANN was able to predict the TSP concentration in Ipatinga atmosphere using meteorological and seasonal cycle data. On balance, the tendency of simulated values was to overestimate the real values. The best modelling results were obtained in BR, BA and CA. BA had the best result and NC, the worst. BA monitoring site is favourable to suffer particulate emissions from Usiminas more than the other sites, as it is located downwind from it in regular meteorological conditions. It could explain why the simulated results were better in BA than in the other ones. The simulated result obtained in NC was the worst due to other variables that were not introduced in the model and its unfavourable location. The type I and II errors results were not representative in the modelling. The type II error only occurred in NC monitoring site and type I error was more common in BA site. Those errors were caused by their location probably. According to statistical analysis of MSE, FB, IA and R, the predicted results from the ANN model were better than the ISCST3 dispersion model.

References

[1] United States Environmental Protection Agency (USEPA). Air Quality Criteria for Particulate Matter – Vol. II – EPA/600/P-99/002a-f, 2004, www.epa.gov/pmresearch/.

[2] Mitkiewicz, G. F., Metodologia para avaliação da dispersão atmosférica de poluentes provenientes de um complexo siderúrgico industrial, 2002, Departamento de Engenharia Sanitária e Ambiental. (Dissertação de Mestrado em Meio Ambiente), Escola de Engenharia da UFMG, Belo Horizonte, Brasil.

[3] Linyan, S. Wang, Y., A neural network model for environmental predication: case study for China. Computers and Industrial Engineering, China, 31, pp. 879-883, 1995.

[4] Wal, J.T., Janssen, L.H.J.M., Analysis of spatial and temporal variations of PM10 concentrations in the Netherlands using Kalman filtering. Atmospheric Environment, 34, pp. 3675-3687, 2000.

[5] Perez, P. Reyes, J., Prediction of particulate air pollution using neural techniques. Neural Computing & Applications, Chile, 10, pp. 165-171, 2001.

[6] Viotti, P.; Liuti, G.; Genova, P. D., Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia. Ecological Modelling, Rome, 143, pp. 27-46, 2002.

[7] Zickus, M., Greig, A.J., Niranjan, M., Comparison of four machine learning methods for predicting PM10 concentrations in Helsinki, Finland. Water, Air, and Soil Pollution, 2, pp. 717-729, 2002.

[8] Perez, P., Reyes, J., Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile. Atmospheric Environment, 36, pp. 4555-4561, 2002.

[9] Podnar, D., Koracin, D., Panorska, A., Application of artificial neural networks to modelling the transport and dispersion of tracers in complex terrain. Atmospheric Environment, 36, pp. 561-570, 2002.

[10] Ordieres, J.B., Vergara, E.P., Capuz, R.S., Salazar, R.E., Neural network prediction model for fine particulate matter (PM2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environmental Modeling & Software, 20, pp. 547-559, 2005.

[11] Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., A neural network forecast for daily average PM10 concentrations in Belgium. Atmospheric Environment, 39, pp. 3279-3289, 2005.

[12] Gardner, W. M.; Dorling, R. S., Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences. Atmospheric Environment, 32(14/15), pp. 2627-2636, 1998.

[13] Haykin, S., Neural networks: a comprehensive foundation, Prentice-Hall Inc.: Canada, pp. 1-842, 1999.

[14] Matlab R12, Version 5.1, The language of technical computing: getting started with Matlab, The Mathworks Inc., pp. 1-86, 1997.

[15] Zhang, G. Patuwo, B.E. HU, M. Y., Forecasting with artificial neural networks: the state of the art. International journal of forecasting, 14, pp. 35-62, 1998.

[16] Kovács, Z.H., Redes neurais artificiais: fundamentos e aplicações, Editora Livraria da Física: São Paulo, pp. 1-174, 2002.

[17] Olesen, H., Model validation kit – status and outlook, National Environmental Research Institute: Denmark, 1997.

[18] Lima, M. M. C., Estimativa de concentração de material particulado em suspensão na atmosfera por meio da modelagem de redes neurais artificiais (Dissertação de Mestrado em Meio Ambiente), Escola de Engenharia da UFMG, Belo Horizonte, Brasil.