COMPARISON OF TWO $PM_{2.5}$ FORECASTING MODELS IN OSORNO, CHILE

PATRICIO PEREZ & BYRON NUÑEZ
Departamento de Fisica, Universidad de Santiago de Chile, Chile

ABSTRACT

According to a recent study of the World Health Organization (WHO), Osorno, a medium sized city in the south of Chile is among the most polluted city in South America. With 150,000 habitants, the city has unfavorable conditions for pollutant dispersion. It is located in a valley between the Andes mountains and a coastal range. Thermal inversions that trap particles emitted mainly by wood combustion used for heating are frequent during fall and winter. Air pollution forecasting models will be useful for authorities to implement a policy of restrictions to emissions when necessary. These models are also useful for the habitants so that they have the possibility to avoid places where air quality is critical, and also so they can choose to restrict physical exercise. The specific meteorological and pollution variables (mostly associated with wood combustion) that can be used as input for statistical $PM_{2.5}$ forecasting models are identified. Results obtained with two models based on artificial neural network techniques: a multilayer perceptron (MLP) and a radial basis function model (RBF) are presented. Both models show comparable accuracy. With them it is possible to anticipate, 24 h in advance, more than 80% of the high concentration episodes during 2018.
Keywords: air quality forecasting, neural networks, air pollution episodes.

1 INTRODUCTION

Osorno is a medium size city located in the south of Chile at 40°57'S and 73°08'W and it is 900 km south of the capital of the country, Santiago. Average altitude is 19 m over sea level and lies in a valley between the Andes mountains and a coastal range. At present, the city has a population approaching 150,000 habitants. Fig. 1 shows a satellite picture of the city in which we can observe the Andes mountains to the right and the coastal range and the Pacific Ocean to the left. Annual average temperature is 11°C, average wind speed is 1.6 m/s and average precipitation is 1,300 mm. Most of this rain is concentrated in the cold season that can be defined by the period between April and September. It is during this cold season that episodes of $PM_{2.5}$ pollution are observed (annual average is 37 $\mu g/m^3$). Most homes use wood stoves for heating and this represents of the order of 90% of $PM_{2.5}$ emissions. Fine particulate matter emitted by stoves is not easily dispersed on dry days because of unfavourable topographic and atmospheric conditions. Low temperatures facilitate the occurrence of thermal inversions due to surface cooling. Under these conditions pollutants become trapped bellow an altitude of no more than 100 m [1]. Daily concentrations often exceed 170 $\mu g/m^3$ which is defined as an Emergency condition. In order to take actions in order to preview these high pollution episodes it seems convenient to have in operation a forecasting model. A statistical forecasting model uses historical data of meteorological and pollution data in order to fix a number of adjustable parameters and rests on the assumption that future behaviour will obey similar functional relation [2]. Most used statistical models for particulate matter forecasting are multi linear regressions (MLR), multilayer perceptrons (MLP) and radial basis function networks (RBF). Stadlober et al. [3] have shown that a MLR model is efficient for the forecasting of daily PM_{10} in three cities in the Alps region. MLP has proven to be a useful tool for NO_2 and PM_{10} forecasting in a populated area in China [4]. Lu et al. [5] report a calculation using a RBF model, and they claim that this method is faster and more effective than more traditional neural network

WIT Transactions on Ecology and the Environment, Vol 236, © 2019 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR190111

models for the forecasting of particulate matter and nitrogen oxides. In this paper the performance of two $PM_{2.5}$ forecasting models based on artificial neural networks, a multilayer network (MLP) and a radial basis function network (RBF), adapted to Osorno conditions is presented. Emphasis is given to forecast high concentrations episodes.

Figure 1: Satellite picture of the city of Osorno.

2 DATA

Data analyzed corresponds to years between 2015 and 2018, with emphasis on the cold period (April–September). Pollution and meteorological information is obtained from an official monitoring station located in the downtown area. According to Chilean pollution legislation, $PM_{2.5}$ 24 h average concentrations are classified into five ranges, which from low to high concern may be labelled in the following manner: range A, corresponding to concentrations bellow 50 $\mu g/m^3$, range B, concentrations between 50 $\mu g/m^3$ and 80 $\mu g/m^3$, range C, concentrations between 80 $\mu g/m^3$ and 110 $\mu g/m^3$, range D, concentrations between 80 $\mu g/m^3$ and 110 $\mu g/m^3$ and range E concentrations greater than 170 $\mu g/m^3$. The Chilean standard for this quantity is 50 $\mu g/m^3$, so only days in range A are considered "safe" days. The Ministry of the Environment has established a plan for the management of pollution in the city which applies for the period between April and September (cold season). During cold season, wood stoves do not properly certified are permanently prohibited. The same

applies for bush burning. For days in ranges D and E, additional restrictions to emissions are enforced. During range D days, industrial and residential boilers identified as responsible of high emissions are not allowed to operate. Also, outdoor physical exercise in schools is prohibited. In the case of range E days, in addition to restrictions applied on D days, the emission of visible smoke from residential heating is banned.

3 MODELING

The goal of this study is to forecast the maximum of the 24 h moving average for the next day (which defines the range of the day) based on pollution data and meteorology available at 19 h of the present day. In this way it is possible to generate an air quality report at 20 h. In order to identify the predictor variables to feed the statistical forecasting model, we did a correlation analysis of candidate variables with the target variable. The selected predictor variables are:

1) Average $PM_{2.5}$ concentration between 1 am and 7 pm of present day.
2) Hourly $PM_{2.5}$ measured at 7 pm of the present day.
3) Hourly PM_{10} measured at 7 pm of the present day.
4) Average PM_{10} concentration between 1 am and 7 pm of present day.
5) Minimum temperature for the next day.
6) Temperature at 8 am of the present day
7) Wind speed at 7 pm of present day.
8) Maximum wind speed for the next day.
9) Maximum relative humidity for the next day.
10) Amplitude of wind direction of present day.

From this list, variables 5, 8 and 9 are forecasted by an independent meteorological model. Variables 1–4 are associated with the tendency of concentrations and are important in the evaluation of 24 h averages on the next day. They are indirectly related to the intensity of wood combustion. Temperature variables are correlated with wood stove usage. Wind related variables are associated with probability of dispersion of particulate matter contained in smoke emitted by stoves. Relative humidity is an indicator of the presence of rain.

Two models based on artificial neural networks are developed. One of them is the traditional Multilayer Perceptron trained with the Backpropagation algorithm [6] (see Fig. 2).

Here, the 10 input variables are connected to units or neurons in the hidden layer by means of weights $w^{(1)}_{ij}$ and activation of these units is calculated by a sigmoid function of a linear combination. Every neuron of the hidden layer is connected to the output by means of weights $w^{(2)}_{k}$ and activation of this unit is calculated again by a sigmoid of a linear combination. Weights calculated during a training phase by an optimization algorithm based on a sample of the available historical data. In our case, 2015, 2016 and 2017 data are used for training. Best results are obtained with a hidden layer with 40 neurons. With weights fixed, an independent test is performed with 2018 data.

The second model is the Basis Radial Function Network (RBF).

The main difference between RBF and MLP is that in RBF the inputs are not equally connected to the hidden units, but only significantly to those that are within a given distance (calculated on the basis of a vector of inputs) from position vectors (that have the dimension of the number of inputs) of these hidden neurons [7], [8]. Output value is calculated through a linear combination of Gaussian functions centered on the position

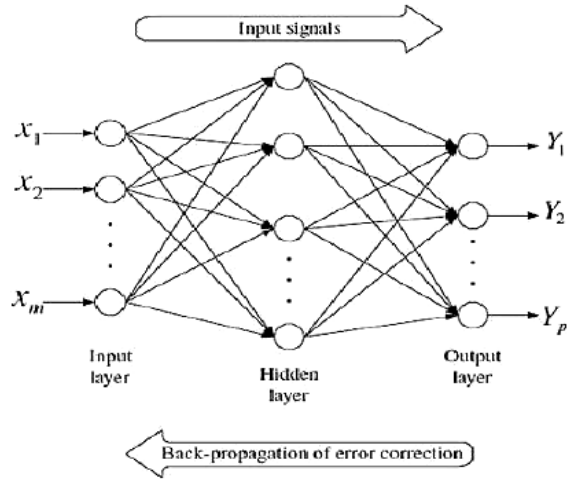

Figure 2: The MLP network.

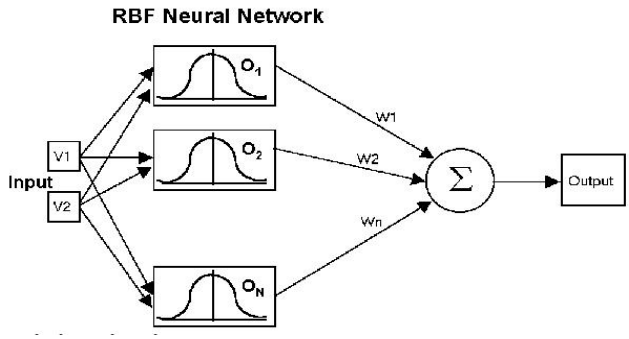

Figure 3: The RBF network.

vectors. Position vectors of the hidden layer are calculated by inspection of the input variable space, generating segregation by clustering. Weights from hidden layer to output are calculated through an optimization algorithm based on a training set. In our case, best results were obtained with 1,000 hidden units. Eventually, RBF may be more accurate than MLP if this clustering is clearly present in the data. In general, with RBF we have a faster training phase and with MLP a faster test phase. MLP has proved to be an efficient algorithm for air pollution forecasting in two Chilean cities, for both daily and hourly averages and for particulate matter and gas pollutants [9]–[12], but RBF has shown to be more accurate than MLP for nitrogen oxide forecasting in Spain [13].

4 RESULTS

The performance of the models is evaluated by testing 2018 data. Indicators of this performance are the mean absolute percentage error (MAPE), root mean squared error (RMSE) [14] and the global model quality observed in contingency tables, which is important to visualize range forecasting.

Table 1 shows the results of the best forecasting MLP and RBF models developed with 2015, 2016 and 2017 data and tested with 2018 cold season Osorno data. It is observed that both models are of comparable accuracy.

Table 1: MAPE and RMSE for 2018 Osorno data using MLP and RBF models.

	MLP	RBF
MAPE	20%	21%
RMSE	39.6	39.5

Although from Figs 4 and 5, an apparently similar performance for both models may be concluded by plotting observed and forecasted values, a more detailed analysis provided by contingency tables (Tables 2 and 3) shows that the MLP model is more accurate on range forecasting. This is more evident for range D, 80% agreement using MLP against 71% with RBF model, and for range C, 48% agreement with MLP against 39% with RBF. Other useful information can be extracted from the contingency tables. As an example, for the MLP model, from the 28 observed range E days during 2018, 23 were correctly forecasted, 3 of them were forecasted as range D and 2 as range C days. It is worth to mention that to get the displayed results with the MLP network, a long time for training and adjustment with different initial conditions was needed. In the case of the RBF network, the fact the best results were obtained with a hidden layer with a number of units of the order of magnitude the size of the training set implies that not significant clustering of the input data was found.

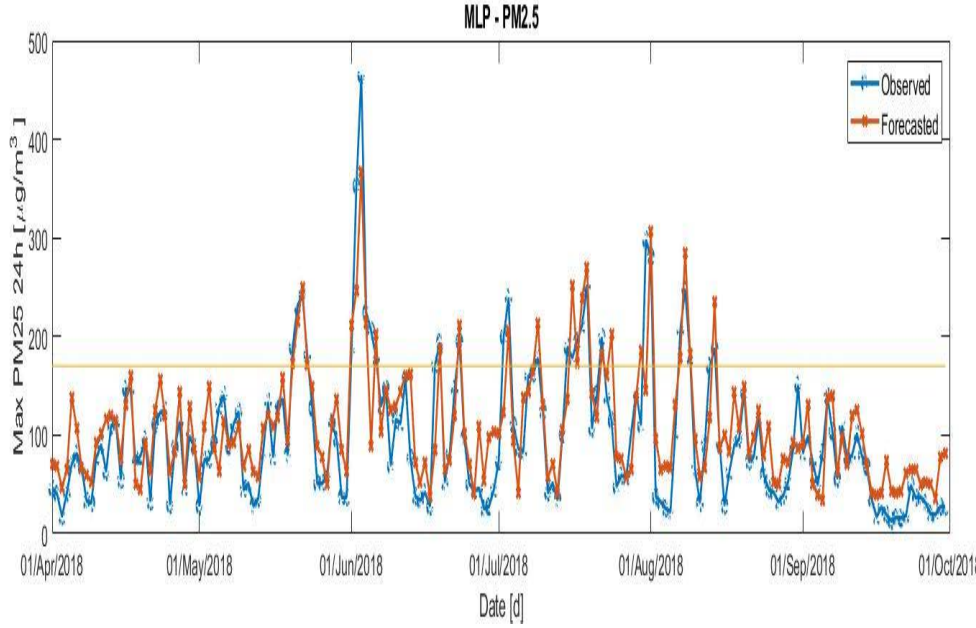

Figure 4: Observed and forecasted maximum of next day $PM_{2.5}$ 24 h average using MLP model. 2018 Osorno test data. Horizontal line indicates level of range E.

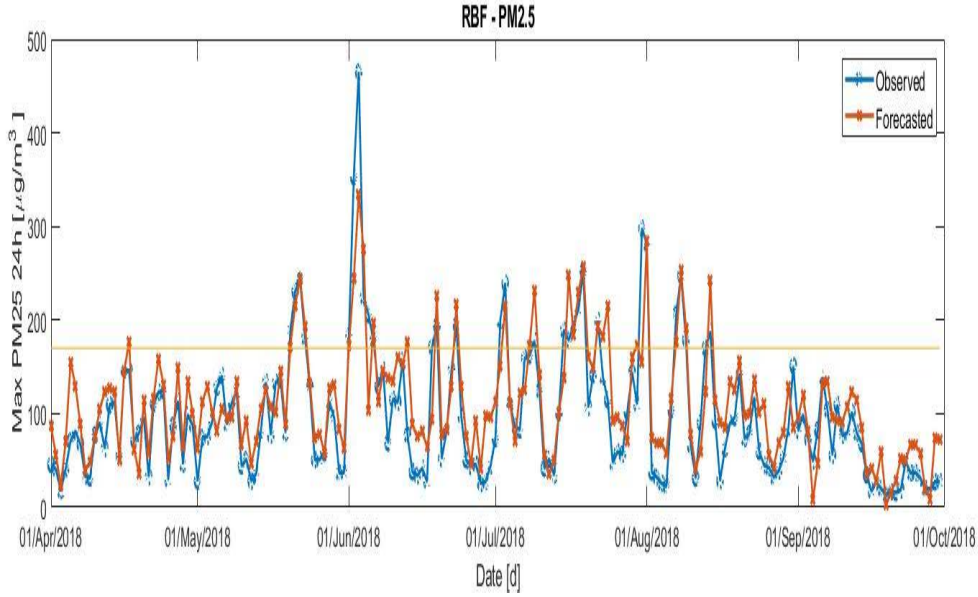

Figure 5: Observed and forecasted maximum of next day $PM_{2.5}$ 24 h average using RBF model. 2018 Osorno test data.

Table 2: Contingency table for observed and forecasted ranges using MLP model.

			Forecast MLP model					
		A	B	C	D	E	TOT	%O
OBS	A	**13**	33	8	0	0	54	24
	B	1	**16**	12	6	0	35	46
	C	2	3	**15**	10	1	31	48
	D	0	1	**5**	**28**	1	35	80
	E	0	0	2	3	**23**	28	82
	TOT	26	53	42	47	25	*183*	*47*
	% F	81	30	35	60	92		

Table 3: Contingency table for observed and forecasted ranges using RBF model.

			Forecast RBF model					
		A	B	C	D	E	TOT	%O
OBS	A	**20**	24	9	1	0	54	37
	B	4	**6**	17	7	1	35	17
	C	1	4	**12**	13	1	31	39
	D	0	1	5	**25**	4	35	71
	E	0	0	2	3	**23**	28	82
	TOT	25	35	45	49	29	*183*	*47*
	% F	80	17	27	51	70		

5 CONCLUSIONS

This study shows that with an appropriate choice of input variables it is possible to implement operational $PM_{2.5}$ statistical forecasting models based on artificial neural network algorithms for the city of Osorno, Chile. The best predictors that apply to this city differ from those used for $PM_{2.5}$ forecasting models in other cities, which may be associated with particularities of this locality. This may help the authorities to take actions in order to protect the population from high levels of pollution. All indicates that shift to heating systems not based on wood combustion would improve air quality significantly.

ACKNOWLEDGEMENT

We would like to thank the support of the Research Department of Universidad de Santiago de Chile (DICYT) through project 091931PJ.

REFERENCES

[1] Molina, C., Toro, A., Morales, R., Manzano, C. & Leiva-Guzman, M., Particulate matter in urban areas of south-central Chile exceeds air quality standards. *Air Qual. Atmos. Health*, **10**, pp. 653–667, 2017.

[2] Shahraiyni, H.T. & Sodoudi, S., Statistical modelling approaches for PM_{10} prediction in urban areas: A review of 21st studies. *Atmosphere*, **7**(2), p. 15, 2016.

[3] Stadlober, E., Hörmann, S. & Pfeiler, B., Quality and performance of a PM_{10} daily forecasting model. *Atmos. Environ.*, **42**, pp. 1098–1109, 2008.

[4] Liu, W. et al., Land use regression models coupled with meteorology to model spatial and temporal variability of NO_2 and PM_{10} in Changsha, China. *Atmos. Environ.*, **116**, pp. 272–280, 2015.

[5] Lu, W.Z., Wang, W.J., Wang, X.K., Yan, S.H. & Lam, J.C., Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, Hong Kong. *Environ. Res.*, **96**, pp. 79–87, 2004.

[6] Rumelhart, D.E., Hinton, G.E. & Williams, R.J., Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, eds D.E. Rumelhart & J.L. McClelland, vol. 1, pp. 318–362, MIT Press: Cambridge, MA, 1986.
[7] Powell, M.J.D., Radial basis functions for multivariable interpolation: A review. *Algorithms for Approximation*, eds J.C. Mason & M.G. Cox, vol. 10, Institute of Mathematics and its Applications Conference Series, pp. 143–167, Oxford University Press: Oxford, 1987.
[8] Moody, J. & Darken, C., Fast learning in networks of locally tuned processing units. *Neural Computation*, **4**, pp. 740–747, 1989.
[9] Perez, P. & Reyes, J., An integrated neural network for PM_{10} forecasting. *Atmos. Environ*., **40**, pp. 2845–2851, 2006.
[10] Perez, P. & Trier, A., Prediction of NO and NO_2 concentrations near a street with heavy traffic in Santiago, Chile. *Atmos. Environ*., **35**, pp. 1783–1789, 2001.
[11] Perez, P., Menares, C. & Ramirez, C., $PM_{2.5}$ forecasting in the most polluted city in South America. *WIT Transactions on Ecology and the Environment,* vol. 230, WIT Press: Southampton and Boston, pp. 199–204, 2018.
[12] Perez, P. & Gramsch, E., Forecasting hourly $PM_{2.5}$ in Santiago de Chile with emphasis on night episodes. *Atmos. Environ*., **124**, pp. 22–27, 2016.
[13] Capilla, C., Application of radial basis functions compared to neural networks to predict air pollution. *WIT Transactions on Ecology and the Environment,* vol. 198, WIT Press: Southampton and Boston, pp. 41–50, 2015.
[14] Botchkarev, A., Performance metrics (error measures) in machine learning regression, forecasting and prognosis: properties and typology, 2018. https://arxiv.org/pdf/1809.03006.