Air quality forecasting in a large city

P. Perez
Departamento de Fisica, Universidad de Santiago de Chile, Chile

Abstract

We describe the different air pollution statistical forecasting models that have been used in Santiago, Chile during the fall/winter period for the last ten years. Effort has been concentrated on particulate matter PM10 for which a standard of 150 μg/m^3 for the 24 h average is currently established. Inputs to the models are concentrations measured at several monitoring stations distributed throughout the city and meteorological information in the region. Outputs are the expected maxima concentrations for the following day at the site of the same monitoring stations. Forecast values using neural network models are compared with the results obtained with linear models and persistence. Recently, a clustering algorithm has appeared as a potentially useful tool to detect high concentration episodes in advance.

Keywords: particulate matter forecasting, neural networks, linear models.

1 Introduction

Air pollution has been a major concern in the metropolitan area of Santiago, the capital of Chile during the last 15 years. Together with Sao Paulo, Mexico City, and some Chinese cities it is considered as one of the most polluted in world. Several factors concur to create unfavorable conditions for air pollutant dispersion. The city is located in a valley that has an extension between 70 and 80 km in the north-south direction and approximately 40 km in the east-west direction. To the west we find the Andes Mountains and to the east a coastal range. Some elevations to the north and south trap the air and air pollutants in a region of poor air circulation, which is enhanced during fall and winter when strong thermal inversions prevent vertical dispersion. During this period of the year, the 24 hour moving average (24MA) of PM10 is used as an indicator of air quality.

WIT Transactions on Ecology and the Environment, Vol 116, © 2008 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR080031

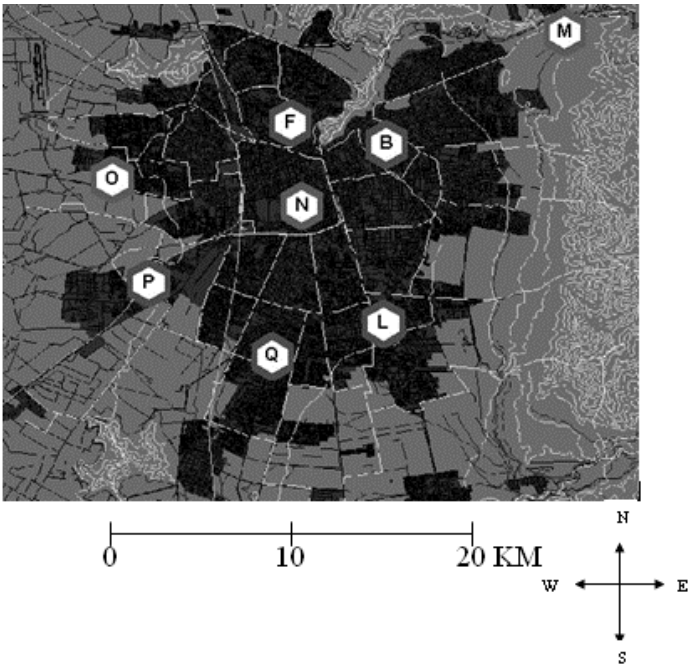

Figure 1: Location of air pollution monitoring stations in the city of Santiago, Chile. The black area represents the urban region. To the extreme right we see the Andes mountains, which spread in the north-south direction.

The main sources of PM10 in Santiago, with six million habitants, are vehicular traffic, industrial activity and heating. Although environmental policies in Santiago during recent years have implied a significant improvement in air quality, particle levels still are considerably high compared to international standards [1]. At present, the standard for the 24 hour average for PM10 in Santiago is 150 $\mu g/m^3$ and the standard for the one year average is 50 $\mu g/m^3$. Regulations that apply on episodes of high concentrations have to do with the definition of classes. According to the value of the maximum of the 24MA of PM10 (MO) the day is classified as class A (good) if MO < 195 $\mu g/m^3$, class B (bad) if 195 $\mu g/m^3 \leq$ MO < 240 $\mu g/m^3$, class C (critical), if 240 $\mu g/m^3 \leq$ MO $\leq$ 330 $\mu g/m^3$ and class D if MO $\geq$ 330 $\mu g/m^3$ .The condition of the city is given by the worst class among all official monitoring stations. On class B days, 40% of the motor vehicles without catalytic converters cannot circulate. On class C days 60% of the vehicles without catalytic converters and 20% of those with catalytic converters are not allowed to circulate. On class C days a number of industries identified as pollutant emitters are enforced to stop operation. On class D days 80% of vehicles without catalytic converters and 40% of those with catalytic converters are not allowed to circulate and more industries are required to stop. Fortunately, the last class D day in Santiago occurred in 2001. It appears very

convenient to have a reliable PM10 forecasting model for the city, which can be used by the authorities in order to warn the population about adverse conditions and to implement palliative actions in advance when extreme conditions are foreseen. In recent years, atmospheric particulate matter forecasting models have been proposed as an aid for air quality management in different parts of the world. Perez et al. [2] have developed neural network, linear and persistence models in order to forecast hourly values of PM2.5 several hours in advance in Santiago, Chile. Three types of neural models, a linear model and a persistence model have been reported in order to forecast the daily averages of PM2.5 in El Paso (USA) and Ciudad Juarez (Mexico) [3]. The performance of multiple linear regressions and neural network models on the forecasting of PM10 in Athens was analyzed by Chaloulakou et al. [4]. Several types of neural network models and a linear model have been used for PM10 forecasting in Helsinki [5]. A multilayer perceptron with emphasis on a novel training algorithm has been used in order to forecast the 24 hour moving average of PM10 in Shanghai [6]. G. Corani has analyzed the performance of neural networks and a linear model locally trained to forecast the daily average of PM10 in Milan [7]. Multilayer neural networks have been used for PM10 forecasting in Santiago since 2002 [8,9]. The results of most of these studies show that neural network models are more accurate than linear models for atmospheric particulate matter concentrations forecasting. A hybrid clustering algorithm (HCA) has also been proposed for forecasting tasks, and it is claimed that it can outperform neural network models [11]. This approach was used for PM10 forecasting, and the results showed a 10% improvement over neural network models [12]. They also agree that sometimes, more important than the particular method, is the appropriate choice of input variables.

2 Forecasting models

The forecasting task may be represented by the implementation of a function of the form:

$$Y = F (x_1, \ldots, x_n, z_1, \ldots, z_m) \quad (1)$$

where Y is a vector with components that are the maxima of tomorrow's 24MA at the site of the monitoring stations, $x_1, \ldots, x_n$ are past values of PM10 concentrations and $z_1 \ldots, z_m$ are measured and forecasted exogenous variables. Input variables may be selected by performing a correlation analysis with historical data.

In the late nineties, restrictions associated to classes B, C, D days in Santiago were applied on the basis of persistence. This means that if on a given day concentrations reached levels within class C, for example, on the following day restrictions associated to that class were applied. This action would make sense only if the episode lasted two or more days.

Since 2001, there is an official forecasting model, which consists of a set of linear equations, one for each monitoring station. The area where most of the times, the highest concentrations are observed is that covered by station O. The equation for this zone is:

$$Y_O = 39.4\ V_O + 0.33\ C_O + 2.06\ T_O + 0.21\ D_O - 21.7 \quad (2)$$

where:

Y_O: is the maximum of the 24 hour moving average of PM10 expected for the following day in $\mu g/m^3$.

V_O: forecasted atmospheric stability for the following day, which is a discrete variable ranging from 1 to 5.

C_O: 24 hour average of PM10 measured at 10:00 AM of present day in station O in $\mu g/m^3$.

T_O: temperature in ºC of the 925 hPa level measured at 12 UTC of present day at a location 80 km west of Santiago.

D_O: change in the last 24 h for height of the 500 hPa level measured at 12 UTC of present day at a location 80 km west of Santiago (in meters).

These last two variables give important information about strength of thermal inversions expected in Santiago in the next hours.

The performance of the forecasting model (worst station) may be evaluated by building a contingency table, which for year 2004 is shown in table 1.

Table 1: Contingency table for official PM10 forecasting model between April 1 and August 16, 2004.

2004		FORECASTED					
		A	B	C	D	TOT	%O
OBSERVED	A	109	15	2	0	126	87
	B	1	6	2	0	9	67
	C	0	1	1	0	2	50
	D	0	0	0	0	0	X
	TOT	110	22	5	0	137	**85**
	%F	99	27	20	X		

In table 1, in columns A, B, C, D we see the number of days forecast to be in a given class against the class of the observed day, which appears in the corresponding arrow. The column %O displays the percentage of observed days by class that were forecast to be in that class. Arrow %F delivers the percentage of forecast days by class that were verified to occur. 100 - %F for each class corresponds to percentage of false forecasts. Numbers in the grey diagonal boxes are the successful forecasts by class. At the lower right corner, the overall rate of successful forecasts is registered. We observe that the performance for this year was reasonable for the identification of class B and class C days, but was poor for the large fraction of false positives on these two classes, which affects the model reliability.

Starting 2003, an alternative PM10 forecasting model for Santiago was presented with the idea to increase the reliability of the instrument on which city authorities base their decisions about restrictions. It was an artificial neural

network model [9]. In this case, Equation (1) is a non linear function that can be schematically represented as a set of nodes connected by weights, in which an input layer contains the variables in parenthesis and the output layer contains the Y components. A hidden layer with a number of auxiliary variables was also included. The transfer function between layers was a sigmoid. Connection weights were calculated by an optimization algorithm that fitted historical data from the previous two years [12]. The inputs used in the neural model were: one hour averages of PM10 measured at 6 PM and 7 PM of the present day at each of five stations (those with highest concentrations in average), the observed difference between maximum and minimum temperature on the present day, the forecasted difference between maximum and minimum temperature on the next day and the forecasted value of an index called PMCA for the next day. This index is a discrete meteorological variable that ranges from 1 to 5 and it is a measure of atmospheric stability in the Santiago area.

Table 2 shows the 2004 contingency table for the neural network PM10 forecasting model.

Table 2: Contingency table for neural PM10 forecasting model between April 1 and August 16, 2004.

2004		FORECASTED					
		A	B	C	D	TOT	% O
OBSERVED	A	119	7	0	0	126	94
	B	2	5	2	0	9	56
	C	0	0	2	0	2	100
	D	0	0	0	0	0	X
	TOT	121	12	4	0	137	**92**
	% F	98	42	50	X		

We observe that the neural model is in overall more accurate than the official model (92% against 85%), it is better for identification of class C days (100% against 50%) and produces less false positives on class B and class C days. Due to change in the properties of emissions in the city, identification of high concentrations (especially class C days) has been poor with both the official model and neural model in the last two years. For year 2007, the contingency table for the neural model is shown in table 3.

This result seems poor considering that the population was exposed to high concentrations of particulate matter when a class C day was verified and no restrictions were applied. It is expected that when the restrictions associated to class C days are applied, they have the effect of lowering to some extent the concentrations. A way to correct the poor performance of the neural model on class C days identification is the proposal by Sfestos and Siriopoulos [10] and Vlachogiannis and Sfestos [11] of a clustering algorithm that may be applied for

Table 3: Contingency table for the neural PM10 forecasting model between April 1 and September 15, 2007.

2007		FORECASTED					
		A	B	C	D	TOT	% O
OBSERVED	A	138	1	0	0	139	99
	B	14	7	0	0	21	33
	C	2	3	2	0	7	29
	D	0	0	0	0	0	-
	TOT	154	11	2	0	167	**88**
	% F	90	64	100	-		

Table 4: Contingency table for the clustering PM10 forecasting model between April 1 and September 15, 2007.

2007		FORECASTED					
		A	B	C	D	TOT	% O
OBSERVED	A	120	14	1	0	135	89
	B	4	13	8	0	25	52
	C	0	1	6	0	7	86
	D	0	0	0	0	0	-
	TOT	124	28	15	0	167	**83**
	% F	97	46	40	-		

air quality forecasting. A natural adaptation of this clustering algorithm has been implemented in Santiago to solve our problem of class identification one day in advance. The algorithm works in the following manner:

For a period of three year training data, we have calculated the average values of the selected input variables within the respective classes (the same variables used in the neural model) A, B, C and D (four centroids). Within every class, we constructed linear or neural networks algorithms that reproduce the values of the output variables (the maxima of 24MA for the sites of the monitoring stations on the following day). Once we have the centroid patterns for each class, we can perform a test with the following year data, by assigning a given vector to the class with centroid to the least Euclidean distance from it. After class identification, we can calculate the numerical forecasted value by using the

algorithm valid for that class. For an operational forecasting system, it would be desirable to generate the most accurate value for tomorrow's class and the expected numerical value of the maximum of the PM10 concentration. The implementation of the clustering algorithm described above with 2007 data produced table 4.

From this table we can verify that the clustering algorithm, having less overall accuracy compared with the neural model (83% against 88%), it has a significantly better performance in detecting class C days (86% against 29%). The false C forecasts would not be so critical considering that most of them were verified to be class B days, which also represent levels considered harmful for the people. A disadvantage of this clustering method is the discontinuity of the numerical forecasted value upon changing from one class to another.

3 Conclusion

With rather simple statistical models it is possible to generate relevant information regarding air quality for the population and authorities in a large city. We have presented several tools that have been used for air quality management in the city of Santiago, Chile and the choice of one of them over the others will depend on the goals we pursue with the forecasting. The models may be used, with the appropriate adaptations in other cities

Acknowledgement

We would like to thank Fondo Nacional de Ciencia y Tecnología (FONDECYT) for support through project 1070139.

References

[1] Koutrakis, P., Sax, S., Sarnat, J., Coull, B., Demokritou, P., Oyola, P., García, J., Gramsch, E., Analysis of PM_{10}, $PM_{2.5}$ and $PM_{2.5-10}$ Concentrations in Santiago, Chile, from 1989 to 2001 *J. Air Waste Manag Assoc* 55, 342–351 (2005).

[2] Perez, P., Trier, A., Reyes, J., Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment* 34, 1189–1196 (2000).

[3] 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 Juarez (Chihuahua). *Environmental Modelling & Software* 20, 547–559 (2005).

[4] Chaloulakou, A., Grivas, G., Spyrellis, N. Neural Network and Multiple Regression Models for PM10 Prediction in Athens: A comparative Assessment. *J. Air Waste Manag Assoc* 53, 1183–1190 (2003).

[5] Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R.,

Cawley, G., Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modeling system and measurements in central Helsinki. *Atmospheric Environment* 37, 4539–4550 (2003).

[6] Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., Shao, D. Progress in developing an ANN model for air pollution index forecast. *Atmospheric Environment* 38, 7055–7064 (2004).

[7] Corani, G. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling* 185, 513–529 (2005).

[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, 4555–4561 (2002).

[9] Perez, P., Reyes, J. An integrated neural network model for PM10 forecasting. *Atmospheric Environment* 40, 2845–2851 (2006).

[10] Sfestos, A., Siriopoulos, C. Time series forecasting with a hybrid clustering scheme and pattern recognition. *IEEE Transactions on systems, man and cybernetics,* Part A 34, 399–405 (2004).

[11] Vlachogiannis, D., Sfestos, A., Time series forecasting of hourly PM10 values: model intercomparison and the development of localized linear approaches. *Air Pollution XIV*, edited by Longhurst, J. W. S. and Brebbia, C. A., WIT Press, 85–94 (2006).

[12] Rumelhart, D. E., Hinton, G. E., Williams, R. J., Learning Internal Representations by Error Propagation. Parallel Distributed Processing. The MIT Press, Cambridge, London, pp 318–364 (1986).