Modelling and Short term Forecasting of Photochemical Pollution by Soft Computing Techniques

G. Finzi

Giuseppe Nunnari

M. Volta

Modelling and Short term Forecasting of Photochemical Pollution by Soft Computing Techniques

Giovanna Finzi[a], Giuseppe Nunnari[b], Marialuisa Volta[a]

[a] *Dipartimento di Elettronica per l'Automazione - Università degli Studi di Brescia - Italy*

[b] *Dipartimento Elettrico Elettronico e Sistemistico - Università degli Studi di Catania – Italy*

Abstract: In this paper, two different approaches, namely of neuro-fuzzy and fuzzy types, are considered for modelling photochemical pollution in two different areas. One of the main aims of the considered approaches is the possibility to extract knowledge from historical time-series thus allowing a deeper understanding of the physical phenomena involved. The city of Brescia is located in the Po Valley in Northern Italy and is characterised by high industrial, urban and traffic emissions and continental climate. The Siracusa industrial area is located in the eastern coast of Sicily, with a climate typical of Southern Mediterranean areas.

Keywords: Neuro-fuzzy networks, fuzzy models, real time alarm system, tropospheric ozone pollution, Decision Support System.

1 INTRODUCTION

Tropospheric ozone (O_3) is a photochemical oxidant, which may cause serious health problems and damage to materials and crops. The European Community directive 92/72/EEC, following the WHO guidelines, prescribes air quality standards for ozone in terms of threshold values for health protection, population information and warning.

The critical anthropic emissions (mainly traffic and combustion processes), the frequent stagnating meteorological conditions and the high solar radiation in Mediterranean regions cause ozone peaks, especially during summer months. In order to take short-term abatement actions and to prevent critical episodes, a proper real time concentration exceedance alarm system has to be set up.

In this paper, two different approaches, namely neuro-fuzzy and fuzzy types are considered for modelling photochemical pollution in two different sites. The city of Brescia is located in the Po Valley in Northern Italy and is characterised by high industrial, urban and traffic emissions and continental climate. The industrial area of Siracusa is located in the eastern coast of the region of Sicily, with a climate typical of Southern Mediterranean areas.

2. THE MODELS

2. 1 The neuro-fuzzy approach

In neuro-fuzzy systems, neural networks are used to tune the *membership functions* of the fuzzy system and to automatically extract *fuzzy rules* from numerical data (Shing et al. 1993). In this work, a four-layer neuro-fuzzy network has been considered (see Figure 1).

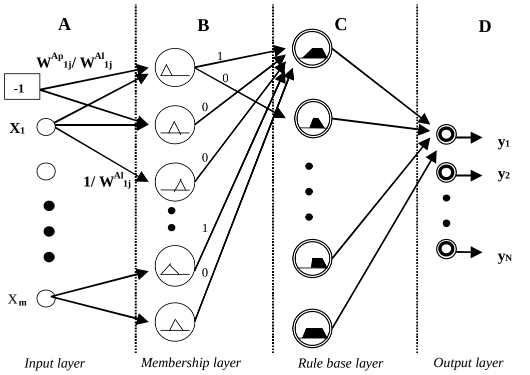

Figure 1. Neuro-fuzzy network architecture.

The nodes of the first layer represent the *crisp* inputs. The activation functions of the second layer nodes act as membership functions. Each neuron of the third layer acts as a *rule node* so that this layer provides the fuzzy rule base. The output of this layer determines the activation level at the output memberships. As ordinary neural nets, the neuro-fuzzy one learns from a training data set, tuning membership functions and rules, by means of a *back-propagation* algorithm. When x_i is the *i*th node in layer *A*, o_j^L is the *j*th output of generic layer *L and* w_{ij}^L is the weight of the link between *j*th neuron at layer *L+1* and *i*th neuron at layer *L*, each layer output can be described as follows (1-3):

Layer B $$o_j^B = \left(1 + \exp\left(-\frac{\left(x_i - w_{ij}^{Ap}\right)}{w_{ij}^{Al}}\right)\right)^{-1} \quad (1)$$

Layer C $$o_j^C = \min_i\left(w_{ij}^B \cdot o_j^B\right) \quad (2)$$

Layer D $$o_j^D = \frac{\sum_i \left(w_{ij}^C \cdot o_i^C\right)}{\sum_i \left(o_i^C\right)} \quad (3)$$

2.2 The fuzzy approach

In the fuzzy approach the prediction problem is formulated in terms of approximating a non-linear time-series *y(t)* in the form of a NARX (Non-linear Auto-Regressive with eXogenus inputs) model:

$$\begin{aligned} y(t+s) = f\,(&y(t), y(t-1),\ldots y(t-n_y+1),\\ &x_1(t), x_1(t-1),\ldots, x_1(t-n_1+1),\ldots,\\ &\ldots, x_q(t), x_q(t-1),\ldots, x_q(t-n_q+1)) \end{aligned} \quad (4)$$

where *f* is an unknown non-linear function, $x_1, \ldots, x_q$ are the exogenous model inputs, *s* represents the number of steps ahead for the prediction model, and n_y, n_1, … n_q are integer numbers related to the model order. Formally, the modelling problem is finding a suitable approximation of the unknown function *f* by using a set of K linguistic rules of the form (5):

$$\begin{aligned} &Ri:\\ &\textit{if } y(t) \textit{ is } A_{i,1} \textit{ and } y(t-1) \textit{ is } A_{i,2}\\ &\quad \textit{and } \ldots y(t-n_y+1) \textit{ is } A_{i,ny} \quad \textit{and}\\ &\quad x_1(t) \textit{ is } A_{i,ny+1} \textit{ and } x_1(t-1) \textit{ is } A_{i,ny+2}\\ &\quad \textit{and } \ldots x_1(t-n_1+1) \textit{ is } A_{i,ny+n1} \quad \textit{and}\\ &\quad \ldots\ldots\\ &\quad x_q(t) \textit{ is } A_{i,ny+n1+\ldots} \textit{ and } x_q(t-1) \textit{ is } A_{i,ny+n1+\ldots}\\ &\quad \textit{and } \ldots x_q(t-n_q+1) \textit{ is } A_{i,p}\\ &\textit{then}\\ &\quad y(t+s) \textit{ is } B_i \quad (i = 1\ldots K) \end{aligned} \quad (5)$$

Where $A_{i,j}$ *(j= 1,…p)* and B_i *(i=1,…K)* are fuzzy sets. In particular, in the case considered here the consequent fuzzy sets B_i are assumed to be *singletons*, i.e. real numbers. The fuzzy modelling approach consists of the following steps:

a) positioning the membership functions $A_{i,j}$ in their respective universe of discourse. This step is based on the determination of the matrix centres of the input data clusters,
b) generation of all possible rules according with the input patterns available,
c) pruning the unnecessary rules. This step is based on approximating the input patterns with the closest cluster centre,
d) determination of the consequent part of each rule. This is done by using a genetic optimisation approach;
e) further pruning phase (this last step is optional) according to a statistical criterion which takes into account the number of each rule activation.

3. PERFORMANCE INDEXES

In order to have a measure of the goodness of the identified models and predictors, the following performance indexes have been defined:

- E[e(t)], the forecasting error expected value;
- σ_e the mean square error,
- σ_e^2/σ_y^2, the ratio between the variance of the error e(t) and the variance of the true time series y(t),
- ρ, the correlation coefficient between true and computed time series.

In order to test the capabilities of the predictors to foresee if the O_3 concentration overcomes an assigned threshold, the European Environment Agency (Van Aalst *et al.* 1997) has defined the following standard *contingency table*:

Table 1. The EEA contingency table.

Alarms	Observed		
Forecasted	Yes	No	Total
Yes	a	f-a	f
No	m-a	N+a-m-f	N-f
Total	m	N-m	N

where:
N is the total number of data points; ***f*** is the total number of forecasted exceedances; ***m*** is the total number of observed exceedances; ***a*** is the number of correctly forecasted exceedances.
Using these definitions, three skill parameters can be defined:

- $\boldsymbol{SP = \left(\frac{a}{m}\right)100\%}$ is the *fraction of correct forecast* smog events (probability of detection) (range from 0 to 100 with a best value of 100). The fraction of *unexpected* events is given by *(100-SP)%*;

- $SR = \left(\frac{a}{f}\right)100\%$ is the *fraction of realised forecast* smog events (range from 0 to 100 with a best value of 100); the *fraction of false alarms* is given by *(100-SR)%*;
- assuming an equal weight to the correct forecasting of smog events and of a non-smog event, the scoring parameters *SP* and *SR* can be combined to a *success index*,

  $$SI = \left(\frac{a}{m} + \frac{N + a - m - f}{N - m} - 1\right)100\% ,$$

  ranging from -100 to 100 with a best value of 100.
- $$S = 100 \cdot \left(1 - \frac{\sum (P(i+1) - M(i+1))^2}{\sum (M(i) - M(i+1))^2}\right),$$

  represents the so-called skill score. In this expression *P(i+1)* and *M(i+1)* represents the predicted and measured value respectively at step *(i+1)*. The main aim of this parameter is to evaluate how much a given prediction model is globally superior to the so-called persistent model. A persistent model is a prediction model defined as follows: P(i+1)=M(i) (i.e. in simple words, tomorrow=today). As it can be easily understood the skill score of a persistent model is zero. Therefore if a given prediction model exibits S >0 it is globally better that the corresponding persistent model.

4. CASE STUDIES

4.1 The Brescia metropolitan area

The examined data records consist of O_3, CO, NO and NO_2 hourly concentrations measured by the urban air quality monitoring station in the city of Brescia. Local temperature monitored and forecasted data are available from the meteorological office. The models are identified on 1994-1998 and validated on 1999 summer season data (from May to September).

The neuro-fuzzy network forecasts are performed at noon on the maximum expected hourly concentration value during the afternoon. The model has been identified assuming triangular membership functions and *sum-prod* inference mechanism. The *crisp* model inputs are O_3 concentrations and the most relevant meteo parameter (temperature) taking part in the photochemical reactions during the day (Finzi *et al.* 2000).

Table 2 shows the inputs and their respective fuzzy set number for the best model as a trade-off between a satisfying forecast performance and a possible operational implementation.

Table 2. The neuro-fuzzy model inputs

Inputs	Value	Fuzzy sets
O3 conc.	10a.m.-12a.m. average	3
O3 gradient	12a.m.-6a.m. difference	4
Temperature	10a.m.-12a.m. average	3
Temperature	12a.m.- 6a.m. difference	2

The rule base turned out to be composed by 30 rules. The *persistent model* skill parameters have also been computed as lower bound performance indexes. The forecast evaluation has been related to an O_3 threshold value of 140 $\mu g/m^3$. Figure 2 compares the skill parameters computed for the persistent and the neuro-fuzzy predictor.

The neuro-fuzzy model seems to be worth using mainly for its skill in avoiding false alarms, while *SP* index claims for a forecast improvement in enhancing some episodes.

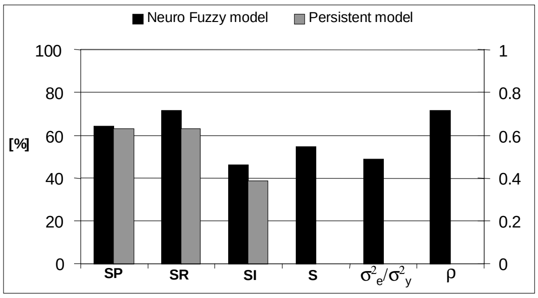

Figure 2. Performance indexes referred to the persistent model

4.2 The Siracusa industrial area

The other area studied is the Siracusa industrial area, in the South-east of Sicily (Italy). In the post-war period, one of the largest concentrations of petrochemical industries in Europe was developed here and it is considered to be an area of high environmental risk.

The air quality in this area is monitored by an interconnected network, run by both public authorities (Provinces and Municipalities) and private organisations. The O_3 time series considered in this paper are those recorded in the station named Melilli. The time series of the O_3 maximum daily values recorded in this station during 1998 and 1999 are reported in the Figure 3.

Fuzzy models have been obtained by using the approach described in section 2.2. Data for model identification were recorded during 1995-1998, while the test was done using data recorded in 1999. The fuzzy model forecasts are performed at 8 p.m. on the maximum expected hourly concentration value during the day after. The inputs of the fuzzy prediction model studied for the Siracusa industrial area are shown in Tab. 3.

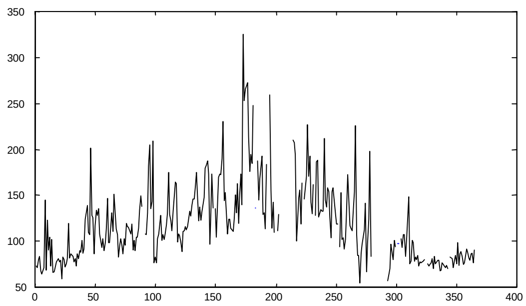

Figure 3a. Melilli,O_3,Max per Day,1998

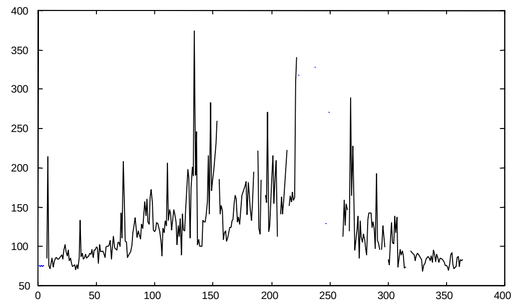

Figure 3b. Melilli,O_3,Max per Day,1999

Table 3. The fuzzy model inputs

Inputs	Value	Fuzzy sets
O_3 conc.	1 a.m. – 8 p.m. average	3
NO_2	1 a.m. – 8 p.m. average	3
NO_X	1 a.m. – 8 p.m. average	3
Temperature	10 a.m. – 6 p.m. average	3
Solar Radiation	10 a.m. – 6 p.m. average	3
Pressure	10 a.m. – 6 p.m. average	3
Wind Direction	10 a.m. – 6 p.m. average	3

The proposed model was compared with a persistent model. The fuzzy model identified consist of 42 rules; 3 fuzzy sets of trapezoidal type for each considered. The results of the comparison carried out with the persistent model for a threshold of 140 $\mu g/m^3$ are reported in Fig. 4.

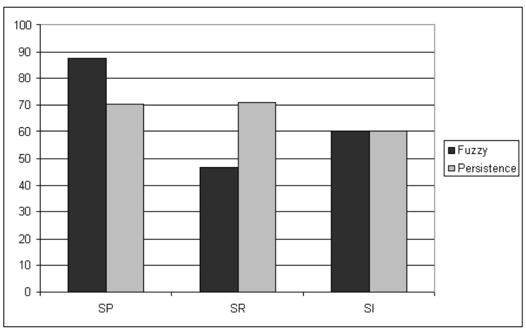

Figure 4. Performance indexes compared with a persistent model for the Siracusa industrial area.

5 RESULTS AND CONCLUSIONS

The neuro-fuzzy model identified for the Brescia metropolitan area, which has been designed specifically to predict critical episodes shows a satisfying performance, both in forecasting over threshold values and in avoiding *false alarms*.

The SP index obtained for the fuzzy model identified for the Siracusa industrial area is considerably better than that exhibited by the persistent model. On the contrary the fuzzy model shows worse performances in terms of SR which results in a larger number of false alarms. This is due to the fact that the performances of the fuzzy model have been optimised with respect to the SP parameters.

Finally it must be observed that:

- fuzzy and neuro-fuzzy predictors perform better than persistent model, i.e. linear ones;
- fuzzy and neuro-fuzzy model complexity and flexibility allow to optimise model performances, stressing the capability to forecast over threshold values or to avoid *false alarms*;
- neuro-fuzzy models, although they are non-linear, can be easily read more than multilayer perceptron neural network and can suggest physical explanation of pollutant processes.

6. ACKNOWLEDGEMENTS

This research has been partially supported by Italian MURST, Brescia University and Provincia di Brescia, in the frame of EUROTRAC-2/SATURN European scientific project. The Siracusa industrial area case study has been carried out in the framework of the EU project APPETISE (Contract N. IST-99-11764)

7. REFERENCES

Finzi, G., and M. Volta, Real time urban ozone alarm system including neuro-fuzzy networks and grey box models, Proc. *Second International ICSC Symposium on NEURAL COMPUTATION-NC2000*, ICSC Academic Press, 659-663, 2000.

Finzi, G., M. Volta, A. Nucifora, G. Nunnari, Real time ozone episode forecast: a comparison between neural network and grey box models, Proc. *International ICSC/IFAC Symposium on NEURAL COMPUTATION-NC'98*, ICSC Academic Press, 854-860, 1998.

Nunnari, G., A. Nucifora, C. Randieri, The application of neural techniques to the modelling of time-series of atmospheric pollution data, *Ecological Modelling*, 111, 187-205, 1998.

Shing, J., and R. Jang, ANFIS: Adaptive-Network-based Fuzzy Inference System, *IEEE trans. on System, Man and Cybernetics*, 23, 665-683, 1993.

Van Aalst, R.M., and F.A.A.M. de Leeuw (eds.), *National Ozone Forecasting System And International Data Exchange In Northwest Europe*, European Topic Centre on Air Quality, 1997.