

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).

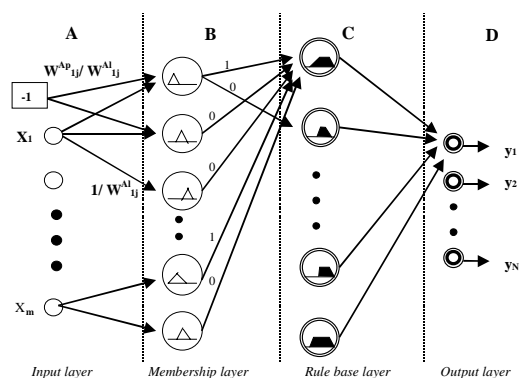


Figure 1. Neuro-fuzzy network architecture.

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):

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

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

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

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 eXogenous inputs) model:

$$y(t+s) = f(y(t), y(t-1), \dots, y(t-n_y+1), x_1(t), x_1(t-1), \dots, x_1(t-n_1+1), \dots, x_q(t), x_q(t-1), \dots, x_q(t-n_q+1)) \quad (4)$$

where f is an unknown non-linear function, x_1, \dots, x_q are the exogenous model inputs, s represents the number of steps ahead for the prediction model, and n_y, n_1, \dots, 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):

R_i :

$$\begin{aligned} &\text{if } y(t) \text{ is } A_{i,1} \text{ and } y(t-1) \text{ is } A_{i,2} \\ &\quad \text{and } \dots y(t-n_y+1) \text{ is } A_{i,n_y} \quad \text{and} \\ &x_1(t) \text{ is } A_{i,n_y+1} \text{ and } x_1(t-1) \text{ is } A_{i,n_y+2} \\ &\quad \text{and } \dots x_1(t-n_1+1) \text{ is } A_{i,n_y+n_1} \quad \text{and} \\ &\dots \\ &x_q(t) \text{ is } A_{i,n_y+n_1+\dots} \text{ and } x_q(t-1) \text{ is } A_{i,n_y+n_1+\dots} \\ &\quad \text{and } \dots x_q(t-n_q+1) \text{ is } A_{i,p} \\ &\text{then} \end{aligned} \quad (5)$$

$$y(t+s) \text{ is } B_i \quad (i=1 \dots K)$$

Where $A_{i,j}$ ($j=1, \dots, p$) and B_i ($i=1, \dots, K$) 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:

- 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,
- generation of all possible rules according with the input patterns available,
- pruning the unnecessary rules. This step is based on approximating the input patterns with the closest cluster centre,
- determination of the consequent part of each rule. This is done by using a genetic optimisation approach;
- 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
	Yes	No	Total
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:

- $SP = \left(\frac{a}{m} \right) 100\%$ 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 $(100-SP)\%$;

