Web: www.witpress.com Email witpress@witpress.com
Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

Neural network based air quality data filling

G. Latini, G. Passerini & S. Tascini
Dipartimento di Energetica, University of Ancona

Abstract

Neural Networks (NN) have become a fundamental tool among data-handling procedures and even more concerning environmental data. In this work we present an application of neural networks to air quality data prediction. Both primary pollutants (mainly SO_2, CO) and photochemical pollutants (particularly ozone) have been considered but the focus has been set on statistical correlation between precursors and secondary pollutants.
After a preliminary study of the phenomena, the work consisted in the following steps: NN architecture choice (we considered Multi-layer Perceptron Networks, recurrent networks and Self Organising Networks), NN set-up, and input handling. Ozone precursors (e.g. NO_x) and meteorological variables have been considered (solar radiation, wind velocity and temperature), noticing that only non-linear relationships were present. We performed an input correlation analysis and we considered normalisation processes and post-training analysis. For the NN training we selected the most representative periods regarding ozone cycle. The final step was the network validation: generalisation capability and prognosis of never processed data-set have been verified.
To maximise the process automation, a software tool has been implemented in the Matlab™ environment. The NN validation showed encouraging results and we successfully extended the SW tool application to the air quality data filling.

1 Introduction

Time series are indispensable elements among the air quality study as well as in many other research fields in which reliable data are necessary for modeling and validation. Series length and completeness are fundamental requirements [1].

Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

Although many advances have been accomplished to realize an extended and reliable monitoring network in the province of Ancona, the actual collected data quality may be surely improved. Intrisic weaknesses of some sensors and inadequate number and location of monitoring stations led to have an unreliable database.

Data losses in a monitoring network may results from several factors belonging to two main categories: instruments failures and output errors.

Typical failures are: Analyser break down, gas sampler malfunction, data acquiring malfunction, power failure, telephone line failure.

Typical errors may be: wrong output reading (e.g. loss of a digit), wrong data logging (e.g. wrong floating point positioning), data-observation swap, wide data gaps.

In order to minimise effects of those inefficences it is necessary to recover as much data as possible. Actual data filling techniques, based on linear and space interpolation and/or statistical regressive models give acceptable results with short data gaps (around 8-10 hours), but wider gaps lead to unacceptable uncertainties on the results.

Recently neural networks (NN) applications among missing data filling are becoming a very amazing alternative to "classical" statistical tools for the capabilities of handling non-linear problems and, more than this, for their generalization capabilities [2] [3]. In this work we present an application of neural networks to air quality data prediction. Although primary pollutant has also been considered (SO_2, CO), the focus has been set on photochemical pollutants, and particularly on ozone, considering statistical correlation between precursors and secondary pollutants.

2 Building up network

First of all it is necessary to split the whole data set into two sub-sets: a training set (examples used for learning, around 70-90% of the whole set) and *test set* (examples used for validation). The idea is to make the network "learn" the training examples and then to test its prediction capabilities. Thus, two main aspects shall be approached: network typology choice and the learning process dimensioning [4].

The NN architecture we opted for is a multi-layer perceptron (MLP) network. In this step, several networks, each one having a different structure (e.g. a different number of neurone in the hidden layer), were set up using the same training set and then all performances have been matched. The best performing network has been considered as the optimal net.

Given a multi-layer network, fixed its complexity, we should find the optimal number of epochs for the best training. An under-trained net will provide poor results and an over-trained net will give excellent result during the training but unreliable ones in the test step.

Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

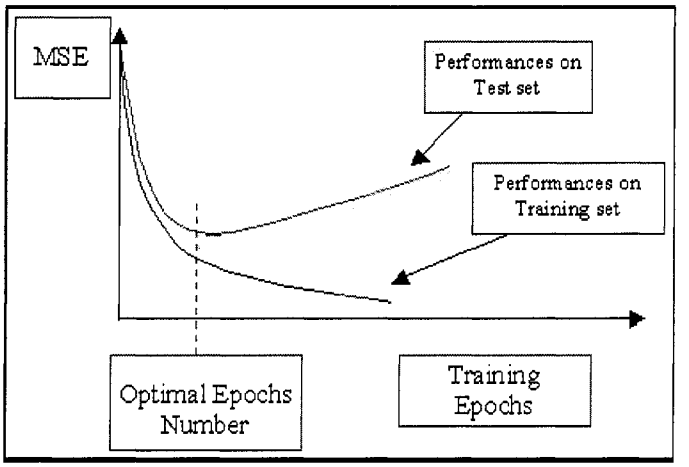

Fig. 1 - Selection of optimal number of epochs

The optimal number of epochs is the one which minimises both MSE and Training error. Since training error decreases when epoch number increases our choise is based on the first parameter (fig. 1). The operative procedure is [5]:

1. make the net learn the training set just for some epochs
2. verify performances on the test set
3. match test set performances with previous ones
4. stop training if error increases

2.1 network dimensioning

There are no confirmed procedure to dimension a multi-layer NN, neither to choose the number of hidden layer nor for perceptron number. Experience is the general way to decide how to set the network up. There are two method to proceed on network dimensioning:

- *network pruning*, starting from an over-dimensioned exceeding branches according to complexity decrease and solution regularity enhancement criteria;
- *network growing*, starting from an under-dimensioned network further neurones will be added to achieve desired performances.

3 The Graphic User Interface (GUI)

Working with NN we found the necessity of implementing a GUI to manage the software tool, allowing a potentially not trained researcher to deal with NN. The package allows a user friendly visualisation of results and makes possible, by means of graphical performance evaluations, the network tuning.

Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

Running the GUI it is possible to choose between trend simulation and data visualisation. After the choice of the air quality quantity to simulate it is possible to select the NN type:

- multi-input dynamic network
- normalised multi-input dynamic network
- preprocessed multi-input dynamic network

Once the desired period has been chosen, it is possible to process, visualise and compare results (fig. 2).

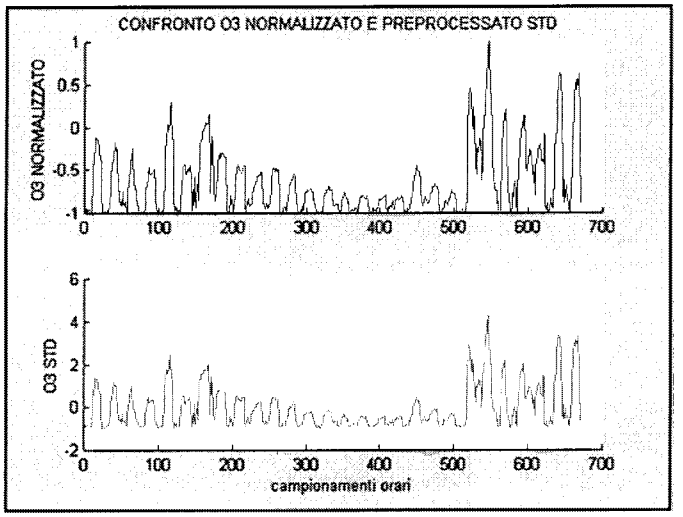

Fig. 2 example of output comparison

4 The Network performances

For this work we considered two stations belonging to Ancona Province monitoring network identified, after their location, "Falconara Scuola" and "Piazza Roma". The first station is located in Ancona suburban area and the second station in the very centre of the town.

The simulations we present in this paper aim to cover the significant periods for Ozone concentration episodes, trying to keep a seasonal representation.

In order to select an adequate working data set we chose time series presenting shortest gaps (no gaps wider than 3 hours) filled up with linear interpolation algorithms. Splitting such data set in a training and a test set we assured the best learning process possible with local data.

As just said above, we implemented multi-input networks, with normalised input and with standard (maximum/minimum, null mean and standard deviation) and p.c.a. (principal component analysis) input pre-processing. Input data, selected for every network, are related to: ozone (pollutant), sulphur dioxide (precursor), wind speed (atmospheric parameter) and solar radiation (atmospheric parameter).

Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

This kind of modelling considers the correlation between a pollutant, its precursors and the solar radiation as positive. On the contrary wind speed has a negative correlation with the pollutant since it tends to increase its dispersion. For every quantity a sensitivity test on net changes of parameters (number of neurones per layer, activation function, training epochs and input delays) has been performed to find the optimal configuration for each condition.

4.1 The basic network

The first network implemented has the structure showed in fig. 3 and presents the following characteristics:

- 4,5 and 1 neurones respectively for the first, second and third layer;
- sigmoidal function for the first layer and linear functions for the others [6];
- 4 regression times (delays) for ozone and 2 for NO_2, solar radiation and wind velocity;
- 40 training epochs.

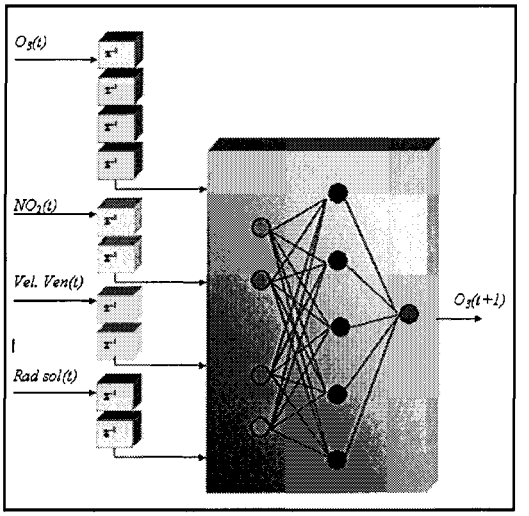

Fig. 3 The basic NN scheme

This network is able to capture quite well the general trend, but uncertainties are evident around peak episodes, tending to under-estimate them.
Along with simulation results, the simulation environment provides several parameters useful to analyse performance and precision: mean error, percentage mean error, percentage standard deviation and validation-step-MSE. While the first parameter is an objective reference the last one is a relative parameter for a performance comparison [7].

Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

Tab. 1 - Performance parameters of the basic network

Series	Mean Error	%ME	% standard deviation	Validation Mse
May 89	5	30	38	42.3
August 94	5	10	9	48.1

4.2 The normalised network

This network has a structure analogue to the "basic" but it also performs an input normalisation It has the following characteristics:

- same number of layers of the previous, but an additional neurone is present in the hidden layer (thus we have 4,6,1 neurones)
- sigmoidal activation factor in the hidden layer since in presence of a normalisation the value range is significantly reduced
- the same delays, since input relations are the same
- training epochs number reduced from 30 to 25 since the network "learns" more quickly with a reduced variation range

The network behaviour is good during training but performances are rather worse than those of the "basic" network.

Tab. 2- Performance parameter of the normalised network

Series	Mean Error	%ME	% standard deviation	Validation Mse
May 89	5	25	30	40.7
August 94	5	17	12	60.5

Performance parameters of the normalised network are reported in tab. 2: a worse behaviour can be noticed for the second series.

4.3 The pre-processed network

The characteristics of this network are:

- 4,6,1 neurones per layer
- sigmoidal activation function for the first two layers and linear for the third one, as for the normalised network
- the same number of delays (4, 2, 2, 2) as for the normalised network
- a lower number of training epochs, namely 20 epochs.

This NN is also characterised by:

- the presence of a pre-processor able to provide null mean and null unit-standard-deviation series
- an analyser of fundamental components able to erase redundant information giving stability to the series

Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

- an hidden layer enhancement (one more neurone with respect to the "basic" network), and so more prediction capability

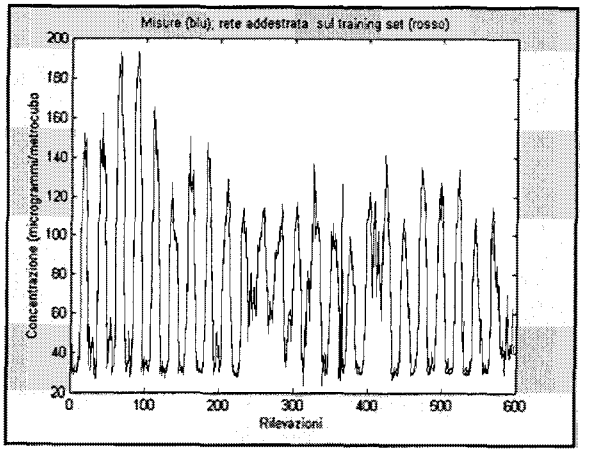

Fig. 4 - training process performance:
observed values overlapped to predicted ones

It is the best network, giving back always the lowest errors and a increased stability. In fact, it has the training performance of the normalised network without the same problems: pre-process not only performs a value scaling but also a statistical re-dimension and a selection of input data.
In table 3, we show results for different periods, related to assorted months, to show the optimal NN behaviour in very different atmospheric condition.

Tab. 3 – Performances of the pre-processed network

Series	Mean Error	%ME	% standard deviation	Validation Mse
August 94	0.6	6	5	16.2
December 98	0.05	5	5	0.6
May 98	1.5	15	16	25.1
September 98	0.2	12	12	20.3

Such good results are visible in fig. 5: data are distributed continuosly along the best fit line resulting in very high correlation factors. Data accumulation around axis origine confirms the limited ozone dynamics in winter season.

Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

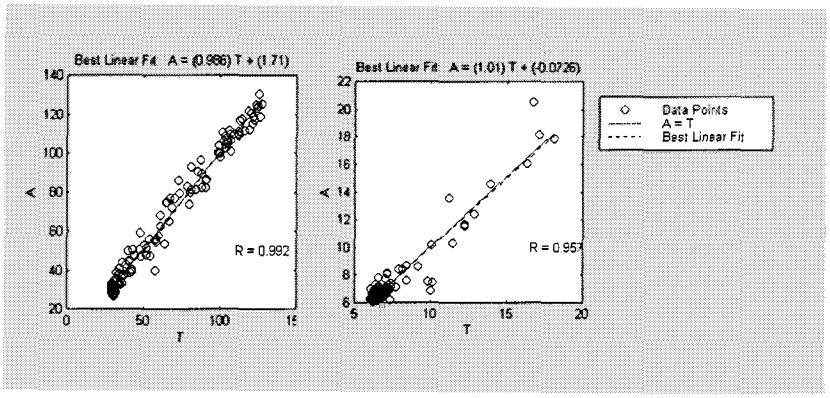

Fig. 5- Training linear regression for August 94 and December 98

A correct error analysis should also consider error dynamics. In fact a good model has a prediction error with no own dynamics: it is said to be *white* with reference to a white spectrum. In other cases data dynamics cannot be captured by model. For this purpose, it is be useful to observe the grphical trend of prediction error auto-correlation. In figures 6 a and 6 b we report the analysis of 36 samples (1 day and a half) related to two typical periods.

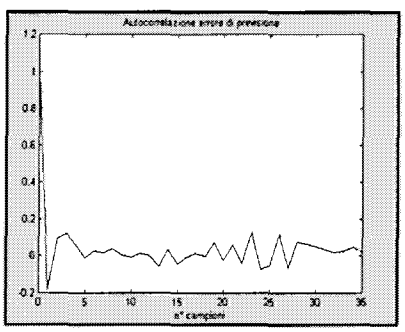

Fig. 6 a) Auto-correlation of prediction error related to August 94

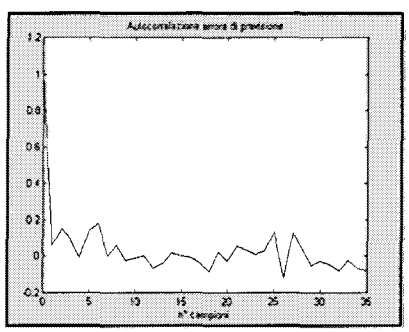

Fig. 6 b) Auto-correlation of prediction error related to December 98

Graphics clearly state that the prediction error of NN output is poorly correlated and it is a further confirmation of the model reliability.

5 The neural network for time series data filling

For time series filling purposes we modified the best performing NN we got (namely the pre-processed network) adding a feedback code. The feedback code

Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

aims to introduce as input values the resulting ones, filling up data gaps, allowing to continue on using the previous ARX modelisation.
The only characteristic that is changed is optimal training epoch number that is raised up to 30. In order to make a comparison possible, we tested performances of the filling NN with the same periods of the previous ones and we imposed the data gap coincident with the previous test set.
Training performance is equivalent to normal pre-processed NN but the test performance is less precise being influenced by the introduction of the predicted values. Anyway, the overall results, depicted in fig. 6 are good.

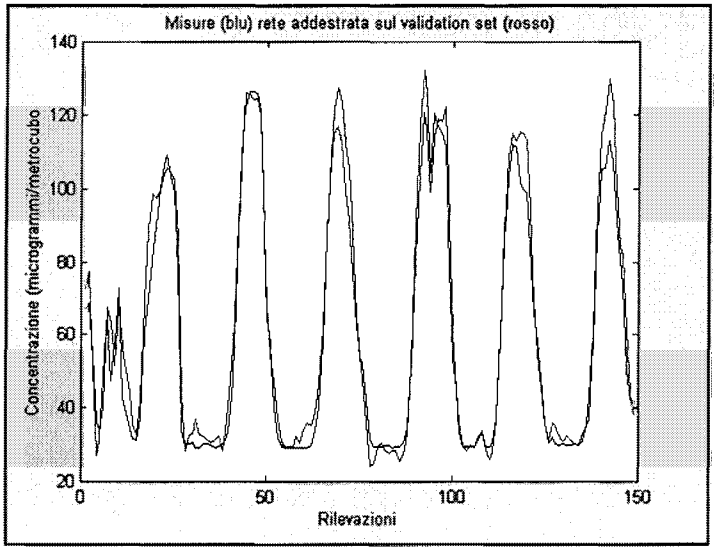

fig. 6 - overlapping of filled values over validation set

Error analysis is good as well: although errors are a little higher, they never exceed 10% and can be considered *white*.

Tab, 4-. Filling NN performance parameters

Mean Error	%ME	% standard deviation
4	11	10

6 Conclusions

A neural network archotecture has been succesfully applied to time series filling problem. We found a network strucutre giving generally good results and the

Paper from: *Air Pollution X*, CA Brebbia & JF Martin-Duque (Editors).
ISBN 1-85312-916-X

filling feedback scarcely affect the NN general performance. Several performance parameters have been considered: standard deviation to MSE, auto-correlation of results and error whiteness analysis all denote that results are encouraging.
We can reasonably consider neural networks as a powerful technique in this field and an increasing application may be foreseen.

Bibliography

[1] R. Stull, "*Boundary layer meteorology"*Kluwer Academic Publishers, 1989.
[2] S. Haykin "*Neural Networks – A comprehensive foundation"* Macmillan College Publishing Company, 1994.
[3] M. Boznar, P. Mlakar. *"Neural Networks – a new mathematical for air pollution modelling."* J. Stefan Institute, Ljubljana, Slovenia
[4] M. Boznar, P. Mlakar. *"A Neural Networks Base Short Term Air Pollution Prediction Model."* J. Stefan Institute, Ljubljana, Slovenia
[5] Rege, Tock. *"A Simple Neural Network for Estimating Emission Rates of Hydrogen Sulfide and Ammonia from Single Point Sources."* Journal of the Air & Waste Management Association. 1996 v 46 n 10
[6] A.C. Comrie *"Comparing Neural Networks and Regresion Models for Ozone Forecasting"* Journal of the Air & Waste Management Association. 1997 v 47
[7] United States Protection Agency (E.P.A.) Office of Air Quality Planning and standards. *Guidelines for developing an ozone forecasting program.* Luglio 1999
[8] Milionis A.E. and T.D. Daevies (1994), ("*Regression and stochastic models for air pollution* "– I. Review, comments and suggestions, Atmospheric Environment, 28,17,pp 2801-2810