BLACK-BOX MODELS FOR PM10 CONCENTRATION FORECAST

M. Milanese
C. Novara
Politecnico di Torino

G. Finzi
M. Volta
Università di Brescia

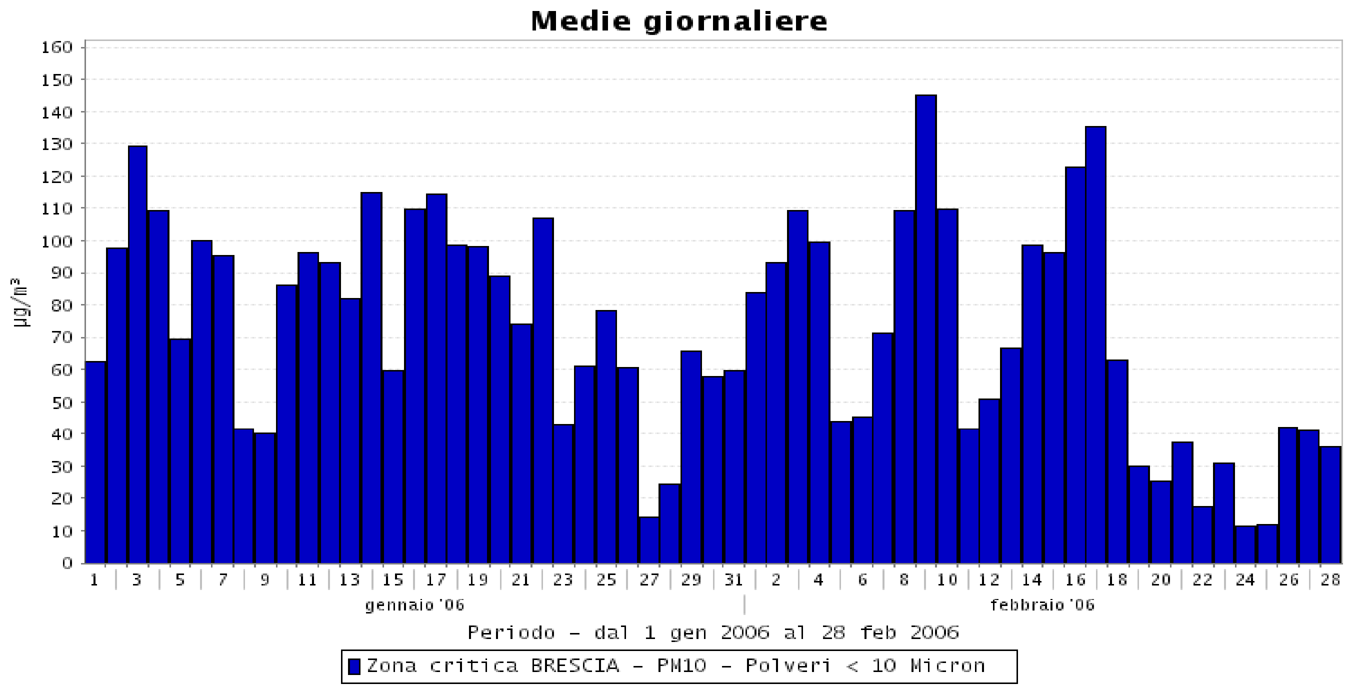
Advanced Atmospheric Aerosol Symposium
Milano, November 12-15, 2006

Outline

- Introduction
- PM10 data set
- Forecast models
- Forecast performances
- Conclusions

Introduction

- Forecast models for PM10 concentrations are important tools for supporting local Authorities in pollution control and prevention.



Introduction

- PM10 processes are complex and non-linear due to the physical-chemical phenomena involving primary particulate and gas precursors (NOx, SO2, NH3 and VOC) in the troposphere.

- Forecast models can be designed formalizing the basic laws, e.g. physical, chemical, economical, biological, etc., of involved phenomena.

- In several applications such as air pollution modelling, the required laws are too complex or hardly known and black-box models are used instead.

- In this paper, forecast of PM10 concentration is approached by means of nonlinear black-box modelling techniques: Neural Network, Neuro-Fuzzy, Nonlinear Set Membership.

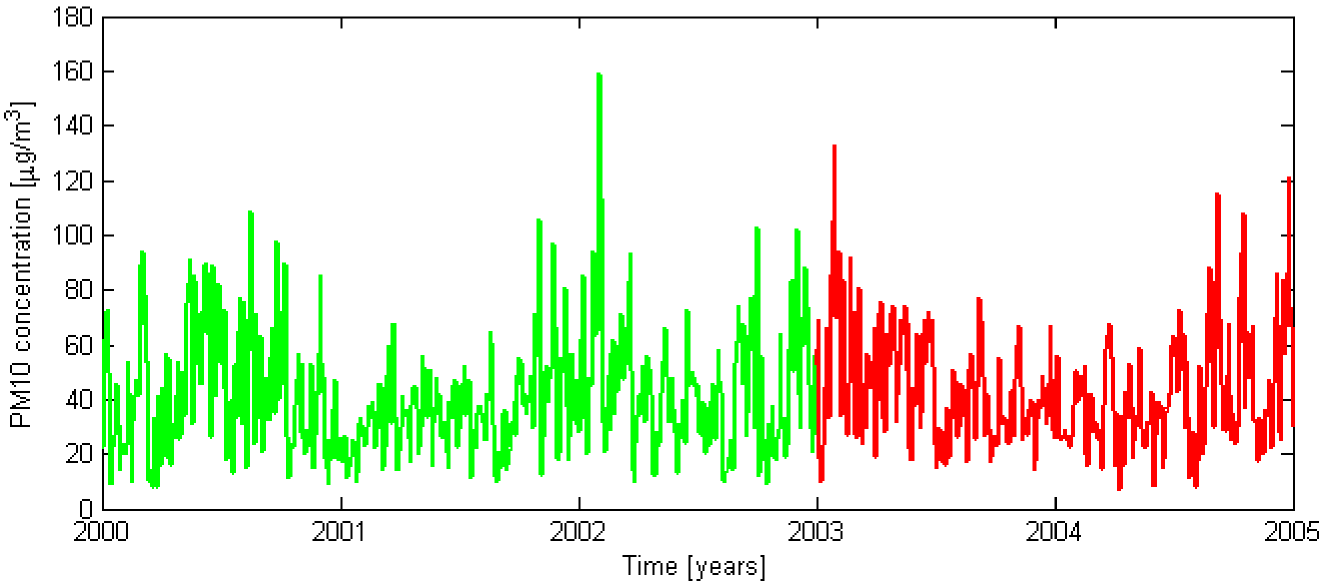
The data set

- Model identification has been performed using data measured in the city of Brescia.
- The city of Brescia is located in the Po Valley in Northern Italy and is characterized by high industrial, urban and traffic emissions and continental climate.
- The examined data records consist of PM10, NOx, CO daily mean concentrations measured by the urban air quality monitoring station during the years 2000-2004.

The data set

- The data set has been partitioned as:

  - ***Identification set***: Years 2000-2002
    Used for model identification.
  - ***Validation set***: Years 2003-2004
    Used for model validation and test.



Forecast models

- The forecast models are of the form:

$$y_{t+1} = f(\varphi_t)$$

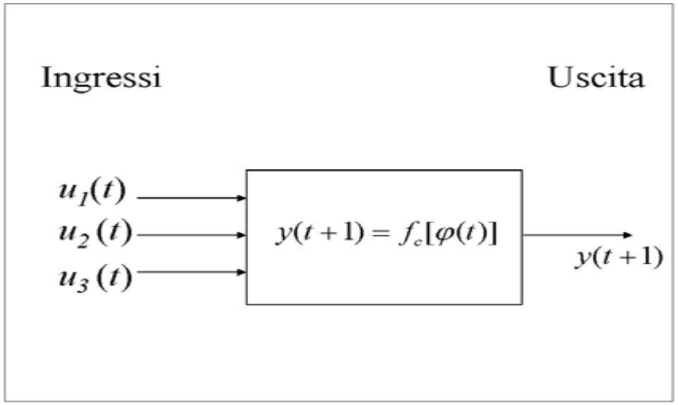
$$\varphi_t = [y_t, u_t^1, u_t^2]$$

t : time step (one day)
y_t : daily mean PM10 concentration at day t
u_t^1 : daily mean NOx concentration at day t
u_t^2 : daily mean CO concentration at day t

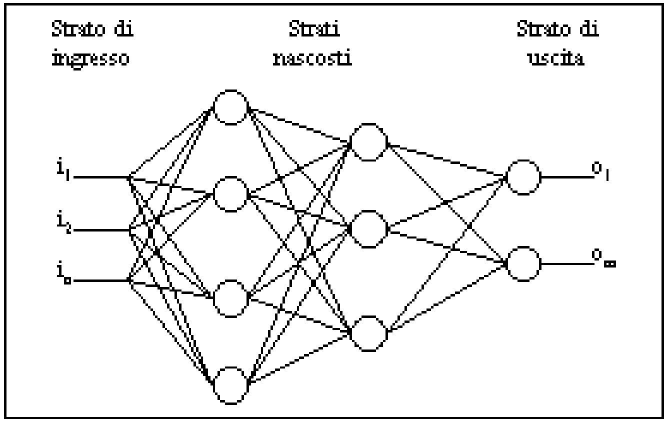
- The model inputs and the lag values have been chosen by means of correlation analysis.

Forecast models

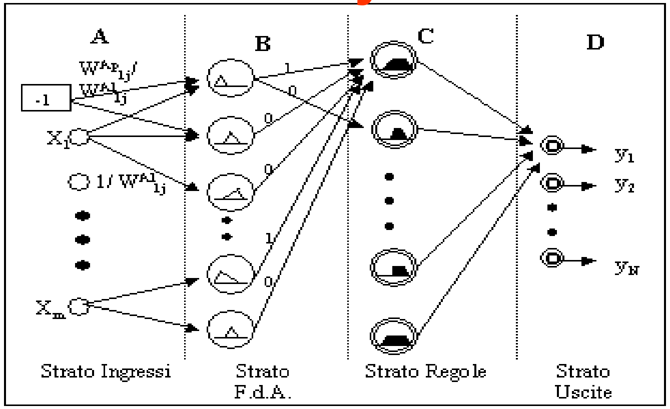
Nonlinear Set Membership



Neural Network



Neuro-Fuzzy



Forecast models

- **Neural Network** model (NN):

$$y(t+1) = \psi[\varphi(t)]$$

$$\psi(\varphi) = \sum_{i=1}^{r} \alpha_i \sigma\left(\varphi^T \beta_i + \lambda_i\right)$$

- $\sigma(x) = 2/(1+2e^{-2x}) - 1$: sigmoidal function
- α, β, λ : net parameters

- Neural networks learn on the training data set, tuning the parameters α, β, λ by means of a back-propagation algorithm.

Forecast models

- **Neuro-Fuzzy** model (NF):

$$y(t+1) = \psi_{NF}[\varphi(t)]$$

ψ_{NF} : neuro-fuzzy function with:
- Membership functions: gaussian
- Inference method: product

- As ordinary neural networks, neuro-fuzzy models learn on the training data set by means of a back-propagation algorithm.

Forecast models

- **Nonlinear Set Membership** model (NSM):

  - No assumptions on the functional form of function f are required. Assumptions on the regularity of f are used:

$$f \in K = \{g \in C^1(\Phi) : \|g'(\varphi)\| \leq \gamma, \forall \varphi \in \Phi\}$$

  This allows to circumvent the complexity/accuracy problems posed by the proper choice of the functional form of f.

  - Noise is assumed bounded:

$$|e_t| \leq \varepsilon_t, \quad t = 0,1,\ldots,N$$

Forecast models

- **Nonlinear Set Membership** model (NSM):

$$y(t+1) = f_c[\varphi(t)]$$

$$f_c(\varphi) = \frac{1}{2}\left[\overline{f}(\varphi) + \underline{f}(\varphi)\right]$$

$$\overline{f}(\varphi) = \min_{t=0,\ldots,N-1}\left(\tilde{y}_{t+1} + \varepsilon_t + \gamma\left\|\varphi - \tilde{\varphi}_t\right\|\right)$$

$$\underline{f}(\varphi) = \max_{t=0,\ldots,N-1}\left(\tilde{y}_{t+1} - \varepsilon_t - \gamma\left\|\varphi - \tilde{\varphi}_t\right\|\right)$$

- **Optimality property**: f_c is the model with minimum worst case identification error among all models which satisfy prior assumptions and are consistent with data.

Forecast performances

- In order to test the capabilities of models to foresee if the predicted concentration will overcome an assigned threshold, the European Environment Agency has defined the following standard 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

N : total number of data points

F : total number of forecasted exceedances

m : total number of observed exceedances

a : number of correctly forecasted exceedances

Forecast performances

- The following performance indexes, defined by means of the contingency table, are used to assess the forecast performances of identified models:
  - SP=(*a*/*m*)100 %: *fraction of correct forecast*
  - *SR*=(*a*/*F*) 100%: *fraction of realized forecast events*
  - *FA*=(100-*SR*)%: *fraction of false alarms*
  - *SI*=[(a/m)+((N+a-m-f)/(N-m))-1] 100%
  - *SK*= $100\,[1-\sum_t\left(\hat{y}_{t+1}-\tilde{y}_{t+1}\right)^2/\sum_t\left(\tilde{y}_t-\tilde{y}_{t+1}\right)^2]$: *skill-score*
  - *CORR*: *correlation between measures and predictions*

Forecast performances

- The forecast performances of the identified models have been evaluated on the validation data set by means of the performance indexes for a threshold of 50 μg/m³

Model	*SI*	*SK*	*SP*	*SR*	*FA*	*CORR*
NN	58,96	19,29	74,38	70,87	29,13	0,62
NF	54,05	16,21	63,64	77	23	0,61
NSM	57,1	19,91	70,5	72,88	27,1	0,63

Green: best result. Orange: worst result.

- **NN**: Neural Network
- **NF**: Neuro Fuzzy
- **NSM**: Nonlinear Set Memebership

Conclusions

- Nonlinear black-box techniques have been applied to the problem of PM10 forecast.
- The identified models provided good prediction performances for the used set of measured variables. In absolute terms, the prediction results are not fully satisfactory.
- All the identified models process the same data and show very similar performances → the information content of the data cannot totally describe the complex PM10 formation and accumulation processes.
- The lack of meteorological (e.g. the wind velocity) and chemical (SO_2, VOC, NH_3 emissions and concentrations) data restricts the skill of the models.