7.08 FIRST RESULTS IN THE PREDICTION OF PARTICULATE MATTER IN THE MILAN AREA

G. Corani and S. Barazzetta
Agenzia Milanese Mobilità Ambiente, V. Beccaria, 19 - 20122 Milano (Italy)

INTRODUCTION

We present the first results of the "Air Sentinel" project, developed at "Agenzia Milanese per la Mobilità e l'Ambiente" and aimed at forecasting pollutant concentrations in the Milan area. This paper describes the firstly accomplished modelling step, which involves the implementation of a couple of linear predictors, able to forecast at 9 a.m. the average PM10 concentration for the current day. The two predictors refer to the two PM10 measuring stations of Milano-Juvara and Milano-Verziere; for the sake of brevity, we will present here the outcome of just the Juvara case. We remark that the presented predictors have been developed by using just data taken from the air quality monitoring network (i.e., pollutant concentrations and basic meteorological data); nevertheless, they provide a satisfactory prediction quality.

In its final stage, the project will forecast pollutants concentrations on the urban area of Milan up to 2 days in advance, by using neural networks predictors. To achieve such a prediction horizon, we will exploit further kind of data, such as meteorological forecasts and radio sondes profiles taken at the surrounding airport of Milano-Linate.

In particular, the instrumentation equipment of the project comprises also a sodar, which has been installed in the very centre of the city to track vertical profiles of micrometeorological variables, such as mean and variances of wind components; sodar measures are furthermore algorithmically processed in order to estimate the mixing layer height. However, such kind of data are currently not included in the predictor, because of the shortness of their time series; nevertheless, they will undoubtedly constitute a precious source of data for improving the forecasts in a near future.

THE CASE STUDY

The Milan urban area, which constitutes the case study of this paper, is located at the centre of the Po Valley, the most industrialised and populated district in Italy. According to the State of the Environment Report [Agenzia Mobilità Ambiente, 2003], the yearly average of pollutants such as SO2, NOx, CO, TSP has decreased respectively of about 90%, 50%, 65%, 60% during the period 1989-2001; no exceedings of alarm and attention thresholds have been observed since 1997 for SO2 and TSP, and since 1999 for CO. The situation is just slightly worse for NOx which, though a clear decreasing trend, showed on the average 8 yearly exceedings of the attention threshold (none of the alarm) over the period 1997-2001 [Agenzia Mobilità Ambiente, 2003]. Also the yearly averages of micro pollutants such as benzene and lead are largely under the thresholds established for human health protection. Such significant results are due to different interventions, such as improved formulations of fossil oils for industrial activities, large adoption of methane for residential heating, and the renewal of circulating motor vehicles.
A severe healthy issue is on the other hand represented by high levels of PM10, which is in practice responsible for all the traffic blocks experienced in the city over the last years. Suspended PM10 origins mainly from the very high vehicular traffic; during winter, also residential heating provides a significant emissive contribution. Moreover, many industries,

power plants, incinerators, factories located in the city outskirts contribute to increase the concentrations. Differently from the above cited pollutants, the yearly average of Pm10 has been substantially stable (about 45 μg/m^3) since the beginning of monitoring in 1998.

The current regional law [DGR 13858 29/7/2003] does not introduce any attention or alarm threshold for PM10; rather, it preventively decrees the blockage of the older pre-Euro vehicles over some specified hour windows during the working days. Nevertheless, we will take into consideration also the capacity of the models in predicting the exceedings of the attention and the alarm thresholds fixed by the previous law [DGR 6501 19/10/2001], respectively set to 50μg/m^3 and 75μg/m^3 . In particular, the law prescribed to declare the *attention state* if the 50μg/m^3 was exceeded for 5 consecutive days. On average, Pm10 exceeds in Milan the attention threshold for about 100 days/year and about 20 attention days have been declared every year [Agenzia Mobilità Ambiente, 2003].

DATA ANALYIS

The data available for the model identification refer to the period 1999-2003 for the Milano Juvara station and to the period 2001-2003 for Milano-Verziere.

PM10 time series underlies a yearly important cyclic patterns, winter concentrations being about as twice as summer ones (Figure 1). Such an effect is due to the combined action of the unfavourable dispersion conditions which are encountered during winter (e.g., reduced mixing layer height) and to the increase in anthropogenic emissions (heavier traffic volumes and build heating) which take place in this season. Therefore, important PM10 pollution events take place almost always during winter. Moreover, the average PM10 weekly profile shows a typical shape, with a decrease of about 25-30% during Sunday with reference to the remaining days of the week; such a pattern is encountered both on the winter and on the summer concentrations data (Figure 2).

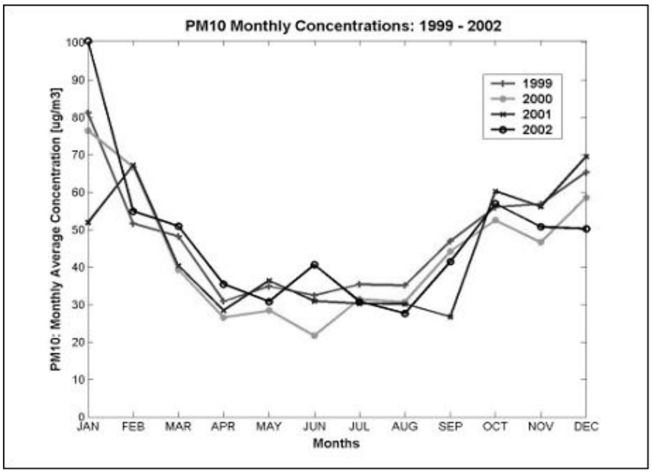

Figure 1. PM10 yearly profile

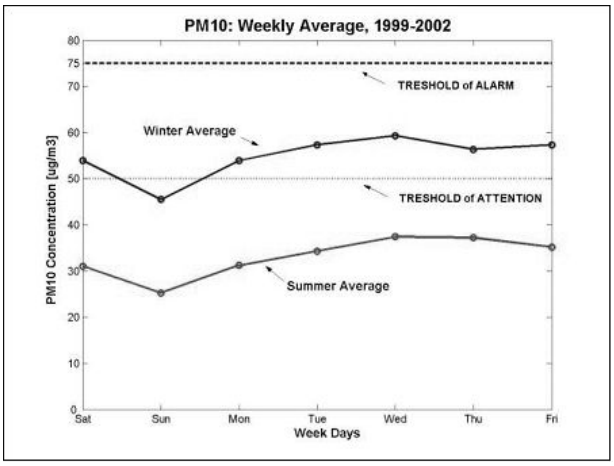

Figure 2. PM10 weekly profile

Variable potentially suitable for the prediction comprise both air quality and meteorological data; to be used as input variable for the predictor, each monitored parameter has to be grouped from the original hourly series to a daily time series. A favourable grouping operator has been identified for each input by means of an exhaustive input/output correlation analysis; in particular, the procedure adopted to identify the optimal averaging window is summarized in the following, where time index (t) refer to a daily step:

1. Set the window begin to 00:00 am of day (t);
2. Set the window end to 01:00 am of day (t);
3. Compute the daily time series of the parameter average on the time window;
4. Correlate the obtained daily time series with the time series of average PM10 of day (t+1), i.e. our prediction target;
5. Increase of one unit the window end, and go back to step 2. Window end is increased up to 07:00 of day (t+1);
6. If the window end reaches 07:00 of day (t+1), reset the window end, and go back to step 2 increasing of one unit the window end. In practice, the procedure is repeated until all the feasible time windows have been investigated.
7. Choose the time window resulting in the maximum correlation with the forecast target

The finally identified set of suitable inputs is made up of the following variables:

PM10(t):
autoregressive PM10, averaged on the time window 22.00 (t) – 8.00 (t+1)

SO2(t):
proxy for building heating, averaged on the time window 13.00 (t) – 5.00 (t+1)

Press. (t):
atmospheric pressure averaged on the time window 1.00 (t) – 7.00 (t+1)

Temp. (t) :
temperature averaged on the time window 4.00 (t) – 8.00 (t)

while the output variable of the model, i.e. the prediction target, is constituted by the PM10 concentrations averaged on the time window 0.00 (t+1) - 23.00(t+1).

Linear modelling approach and prediction performances assessment

The developed model is a linear regressor, mapping a linear combination of the input variables on the prediction target, as shown in the following equation:

$$PM10(t+1) = \alpha\, PM10(t) + \beta\, SO2(t) + \gamma\, Press\ (t) + \delta\, Temp\ (t) \quad (1)$$

Given the linearity of equation (1), the predictor parameters can be estimated in an analytical way through linear least squares, and the identification procedure is especially fast: indeed, parameters are estimated almost instantaneously on a standard PC, even the time series collects more than one thousand time steps.

The prediction performances have been assessed via *k-fold cross-validation*, which is resampling methodology already adopted in important air pollution time series studies (Dorling et al, 2003).
The performances assessment procedure we implemented is as follows:

1. Split the dataset (made up of 57 months in the Juvara case) into a *training set* (54 months) and a *testing set* (the remaining 3 months);

2. Calibrate the linear regressors of equation (1) on the training set;

3. Evaluate the prediction performances on the testing set;

4. Return to step 1 configuring a new couple of training and testing set, until all the months in the dataset have been inserted once in the testing set.

In order to assess the forecast accuracy, the following indicators have been adopted:

- The true/predicted correlation
- The mean absolute error (*MAE*): $MAE = \frac{1}{n} * \sum |y - \hat{y}|$
- The mean bias error (MBE): $MBE = \frac{1}{n} * \sum (y - \hat{y})$
- The probability *P* of detecting a "attention state" declaration
- The probability *FA* of giving a false "attention state" alarm

Table 1. K-fold Cross validation performances

Testing Set/ performances	**1-3 1999**	**4-6 1999**	**7-9 1999**	**10-12 1999**	**1-3 2000**	**4-6 2000**	**7-9 2000**	**10-12 2000**	**1-3 2001**	**4-6 2001**
ρ	0.9169	0.8651	0.8975	0.8774	0.8912	0.9063	0.9127	0.8274	0.8005	0.8464
Sqm	13.35	5.55	7.47	19.08	12.37	5.14	8.36	15.70	14.73	5.70
MAE	9.26	4.40	5.97	14.78	9.10	4.63	6.33	12.45	9.94	4.45
MBE	0.79	0.44	1.89	0.05	-2.43	-1.86	1.67	0.34	0.48	0.30
P	100%	N.A.	83%	82%	83%	N.A.	0%	86%	60%	N.A.
FA	0%	N.A.	0%	9%	17%	N.A.	0%	14%	40%	N.A.

Testing Set/ performances	**7-9 2001**	**10-12 2001**	**1-3 2002**	**4-6 2002**	**7-9 2002**	**10-12 2002**	**1-3 2003**	**4-6 2003**	**7-9 2003**
ρ	0.8933	0.8735	0.8777	0.8697	0.8786	0.8358	0.8347	0.7870	0.9265
Sqm	9.91	19.00	9.66	7.07	11.59	15.44	10.46	6.54	8.76
MAE	7.32	14.63	7.33	5.32	8.57	11.56	7.72	5.23	6.86
MBE	-1.28	1.76	-2.46	-0.000	2.34	0.09	-1.55	0.80	-2.40
P	100%	85%	100%	N.A.	89%	46%	75%	N.A.	67%
FA	0%	15%	0%	N.A.	11%	36%	25%	N.A.	33%

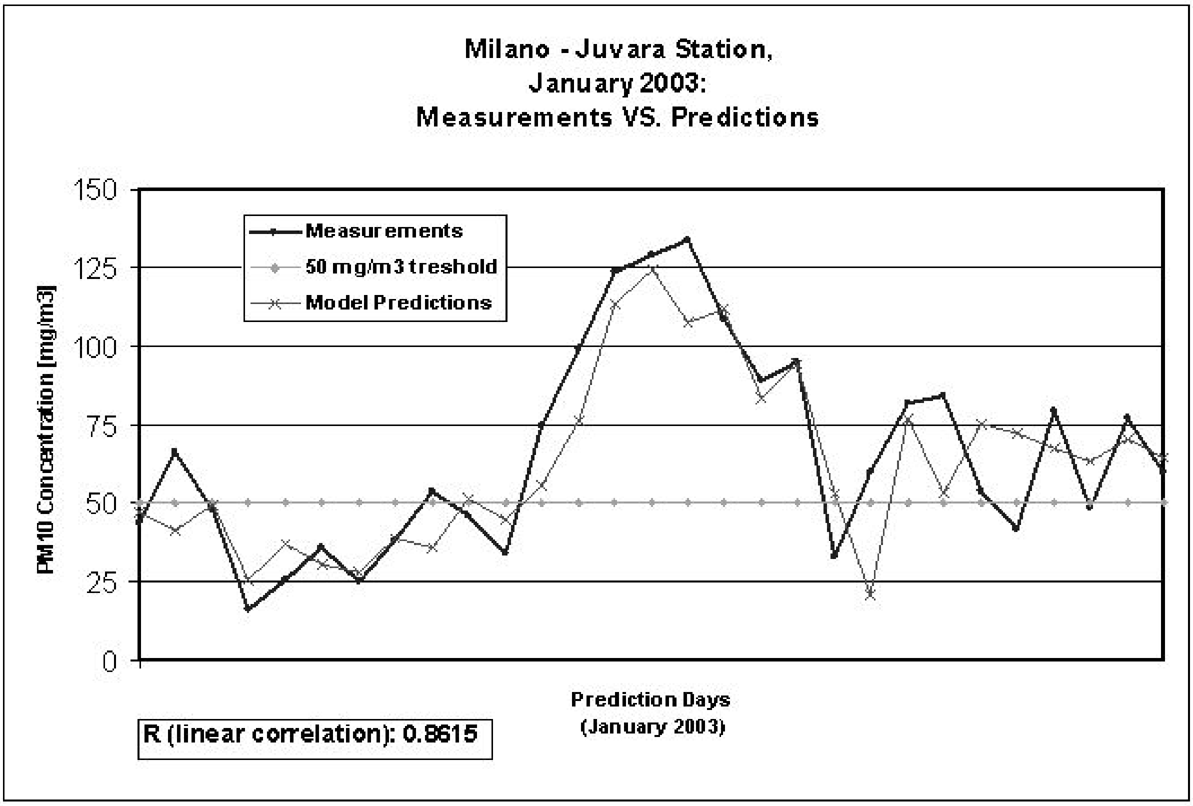

Figure 3. January 2003 simulation

A sample of operational use of the model, regarding the January 2003, is given in Figure 3. The model captures satisfactory the pollution dynamic, even if PM10 exceeds frequently the attention threshold.

CONCLUSIONS

The presented model is able to satisfactory forecast the PM10 concentrations of the current day. However, a great research effort will be devoted in the future or improving the prediction accuracy and exceeding the forecast horizon. Significant improvements are expected from the exploitation of advanced meteorological data, such as sodar, radiosondes, meteorological forecasts.

REFERENCES

Dorling, S.R. and Foxall,R.J. and Mandic,D.P. and Cawley,G.C., 2003: Maximum likelihood cost functions for neural network models of air quality data, *Atmospheric Environment*, **37**, 3435-3443

Agenzia Milanese Mobilità Ambiente 2003: Relazione sullo stato dell'ambiente del comune di Milano