

# PM<sub>10</sub> forecasting with a local linear approach

M. Bianchi<sup>◦</sup>, G. Corani<sup>\*◦</sup>, G. Guariso<sup>\*</sup>

<sup>◦</sup>AMA, Agenzia Milanese Mobilità Ambiente,

<sup>\*</sup>Dip. Elettronica e Informazione, Politecnico di Milano

[mauro.bianchi@ama-mi.it](mailto:mauro.bianchi@ama-mi.it); [corani.guariso@elet.polimi.it](mailto:corani.guariso@elet.polimi.it)

Local linear models are developed and tested in this paper to predict the next day PM10 concentrations in two urban sites in Lombardy. The fast implementation of the algorithm, developed by IRIDIA, Brussels, and called *lazy learning*, is exploited to analyse several alternative combinations of meteorological and air quality input variables and to determine their relative importance.

## 1. Introduction

Most state-of-the-art modelling approaches, both physically based and black-box, such as feed-forward neural networks, are based on a unique relation between a number of input variables and the required output, namely the predicted concentration of a pollutant at a given point. However, such relations are difficult to assess and, apart from the common perception that they are mainly non linear, it is unclear which input variables should be included and how. Furthermore, once such a global model is identified on a training data set, it is normally kept unchanged for all the predictions. In this way no further improvement is possible except for the uncommon case of the recursive update of linear parameters.

An alternative approach is local modelling which renounces to a complete description of the input-output relationship, focusing on approximating the system only in the neighborhood of the point to be predicted. Lazy learning is a local linear modelling approach, developed by the research group at IRIDIA, Free University of Brussels, which performs a complex learning procedure, resulting in the identification of a local model, only when a prediction is required. The identified local predictor is then discarded after having returned the forecast.

The algorithm works in four basic steps. The first is the evaluation of the distance between the current values of input variables (query-point) and all the similar situations (neighbors) available in the identification dataset. Second, neighbors are ranked according to their distance from the current query-point. Third, different models are identified on nearest neighbor sets of different cardinality. Fourth, the model showing the lowest average leave-one-out error (error computed on a single neighbor when the parameters are identified on the others) is finally chosen, and the prediction is computed.

The possibility of identifying the model on neighbor sets of different cardinality allows the algorithm to find the best compromise between an accurate and reliable parameter estimate and the local nature of the model. This obviously implies that the entire identification dataset must always be kept in memory. While this is a drawback in terms

of computational requirements, it implies that the predictor can be kept up to date by simply adding new samples to the identification dataset. This also means that the algorithm can closely follow not only non linear phenomena, but also time varying ones.

The fast lazy learning implementation by IRIDIA is exploited in this paper to develop and test several alternative combinations of input variables for short-term prediction of  $PM_{10}$  concentrations in two urban sites in Lombardy.

## 2. $PM_{10}$ in the Milan area

As stated in the “2005 Air Quality Report” by Agenzia Milanese Mobilità e Ambiente - AMA (Agency for Mobility and the Environment, Milan), the concentrations of pollutants such as  $SO_2$ ,  $NO_x$ , CO, TSP have decreased strongly in the last 15 years; on the other hand,  $PM_{10}$  represents the major concern for air quality in Milan. For instance, the attention threshold of  $50 \mu g/m^3$ , fixed by Ministerial Decree n. 60 April 2, 2002, has been exceeded for about 100 days per year (see Fig. 1) since the beginning of  $PM_{10}$  measurement in Milan (1998), while EU directives limit the maximum number of days of exceedance of such threshold to 35 days per year.

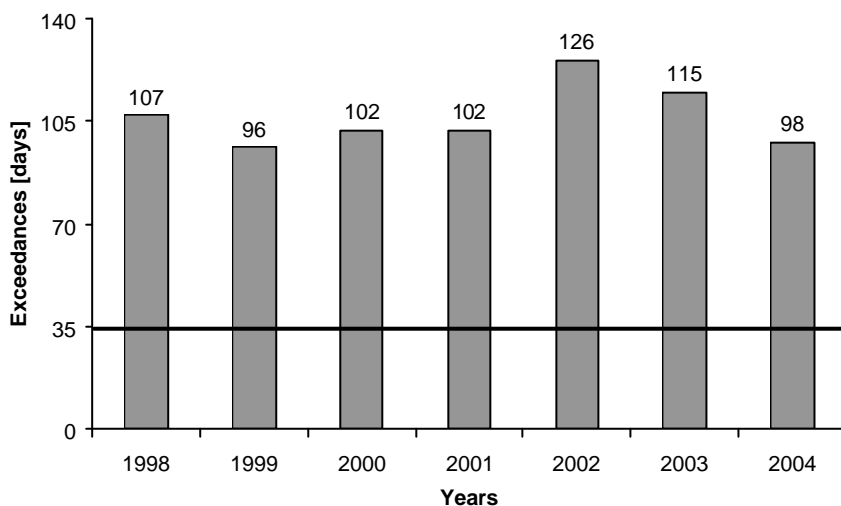


Figure 1: yearly exceedance of the  $50 \mu g/m^3$  threshold, for the Milan-Juvara station. The black line represents the limit of 35 days/year imposed by present EU directives.

This situation is due to many different factors: the Milan urban area presents heavy industrialization, high traffic fluxes and dense population. Moreover, the local climate is characterized by stable atmospheric conditions for most of the year, which prevents particulate dispersion.

The “Air Sentinel” project, developed by AMA aims at publishing daily forecasts of pollutant concentrations in Milan. In this paper we present the results obtained with one-day-ahead and two-days-ahead  $PM_{10}$  predictors.

In particular, we consider two measuring stations: one located in the urban area of the city (Milan-Juvara), and a second one located at the east border of the city (Limito). Models are trained to predict the daily average concentrations of  $PM_{10}$ . Table 1 shows the periods of data availability and some basic statistics about the  $PM_{10}$  time series.  $PM_{10}$  measures were collected by the Regional Agency for Environment Protection (ARPA) using a TEOM (Tapered Element Oscillating Monitor) analyzer, with a 2 hours sampling resolution. Data are publicly available on the website [www.arpalombardia.it](http://www.arpalombardia.it). It is worth to notice that TEOM analyzers, are actually being replaced and new measurement methods (such as gravimetric analyzers) lead to  $PM_{10}$  values that are about 20% higher.

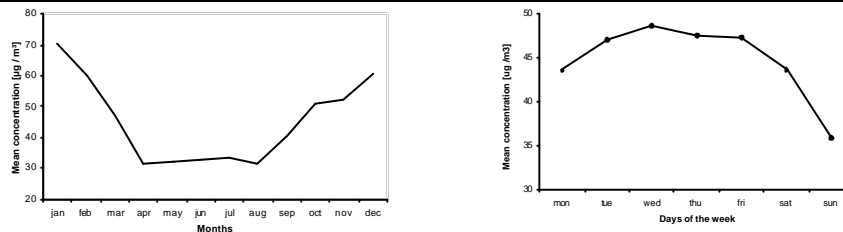
Table 1:  $PM_{10}$  data availability and characteristics

| Station      | Data availability (years) | Mean [ $\mu\text{g}/\text{m}^3$ ] | Standard deviation [ $\mu\text{g}/\text{m}^3$ ] |
|--------------|---------------------------|-----------------------------------|-------------------------------------------------|
| Milan-Juvara | 1999 to 2004              | 43.9                              | 25.1                                            |
| Limito       | 1999 to 2004              | 40.7                              | 25.4                                            |

The time series of  $PM_{10}$  daily concentrations underlies different periodic components, as shown in Fig. 2 for the Milan-Juvara station:

- *yearly periodicity*: winter concentrations are about twice as summer ones, because of both higher anthropic emissions (domestic heating and traffic) and unfavorable dispersion conditions (Fig. 2a);
- *weekly periodicity*: concentrations are lower (25-30%) during weekends than during weekdays, probably because of the reduced traffic fluxes (Fig. 2b).

Similar trends are observed also for the Limito station.



(a): mean yearly pattern

(b): mean weekly pattern

Figure 2: Mean monthly and daily  $PM_{10}$  concentrations observed at the Milan-Juvara station for the period 1999-2004.

### 3 Forecast methodology

In this study a local learning algorithm, known as *lazy learning* and proposed by Birattari et al. (1999), is used to predict  $PM_{10}$  concentrations. Lazy learning has been indeed shown to be a viable approach for air quality prediction (Corani, 2005).

As already pointed out, the basic idea in local modelling is to renounce to a global description of the system under study, focusing on its behavior only in the neighborhood of the current condition. In particular, the lazy learning algorithm is a local linear approach that defers the processing of the dataset until a request for

prediction is received; when this happens, an identification procedure takes place and a local model is designed.

For every request of prediction, the algorithm evaluates the distance between the current query-point  $\mathbf{Q}$  (a vector of  $n$  variables denoted by  $Q_i$ ) and the available historical samples (vectors  $\mathbf{X}$ , whose components are the variables values  $X_i$ ) using the Manhattan metric  $D(\mathbf{Q}, \mathbf{X}) = \sum |Q_i - X_i|$ , computed on standardized data to avoid issues with the measurement units of the different variables. All the training samples are then ranked according to their distance from the current query-point.

Then, different candidate local models are identified using a number of neighbors (i.e., historical samples closest to the current query-point) increasing from the constant value  $k_{min}$  to the constant value  $k_{max}$ ; hence,  $(k_{max} - k_{min} + 1)$  candidate models are identified for each query-point.

The final step involves choosing the best model, which can be seen as choosing the optimal number  $k$  of neighbors to calibrate the local linear model. This requires to address a trade-off, since smaller  $k$  possibly lead to large variance of the parameter estimates, but larger  $k$  may entail a poor representation of non-linearities, since the model tends to a global linear model. The criterion adopted to tune the number of neighbors on a query-by-query basis is leave-one-out cross-validation (Birattari et al., 1999): the dataset corresponding to the current value  $k$  is split into a training subset of cardinality  $(k - 1)$  and a validation subset, containing just the remaining sample. Parameters are estimated on the training set, while the error is measured on the validation sample (leave-one-out error) and then averaged over the complete permutation. Iterating the procedure on all candidate models, a vector of  $k$  leave-one-out errors is generated. The model showing the lowest local leave-one-out error is finally chosen.

In this work, we use the open-source lazy learning package, publicly available at the website <http://www.irisia.ulb.ac.be/~lazy..> It provides a very efficient way to accomplish both tasks of local linear model identification and leave-one-out cross-validation, making use of the recursive least squares algorithm.

### 3.1 Input variables selection

We denote with  $t$  the current day (when prediction is issued), while days  $t+1$  and  $t+2$  are respectively the days for which the one day and two days ahead predictions are evaluated.

Input variables selection is usually performed considering variables that show a high linear correlation with the output and a low inter-correlation within the input set. However, such an approach, based on linear relations between variables, poorly fits with the known non linear link between  $PM_{10}$ , its precursors and meteo conditions. On the contrary, the ease and speed of the lazy learning algorithm allows to test a very large number of input variables combinations to determine the most effective one.

Available input variables are: hourly measures of pollutants and meteorological observations (such as temperature, air pressure, air humidity, solar radiation, wind speed and directions, rain, etc.) coming from the air quality monitoring network and available up to 12 p.m. of day  $t$ ; daily values for the height of the mixing layer, computed by a micro-meteorological model are also available up to day  $t$ . To give a description of the

anthropic component, averaged traffic fluxes on a day of the week basis and the type of day (working day, Saturday, holiday) were also considered. These last two input variables contain very general and less accurate information, but they are clearly available for days  $t+1$  and  $t+2$ .

### 3.2 Performance evaluation

In order to assess the goodness of predictions, we use the same set of performance indicators quoted in Schlink et al. (2003).

Such indicators are the true/predicted correlation  $r$ ,

the mean absolute error  $MAE = (1/N) \cdot (\sum |y(t) - \hat{y}(t)|)$ ,

the mean bias error  $MBE = (1/N) \cdot (\sum \hat{y}(t) - y(t))$ , and

the index of agreement  $d = 1 - (\sum (y(t) - \hat{y}(t))^2 / \sum (|\hat{y}(t) - \underline{y}(t)| + |y(t) - \underline{y}(t)|)^2)$ .

In all these formulas,  $N$  denotes the total number of days,  $\hat{y}(t)$ ,  $y(t)$  and  $\underline{y}(t)$ , the predicted real, and average values for day  $t$ .

We assess then the ability of the model in predicting the exceedances of the 50  $\mu\text{g}/\text{m}^3$  threshold. Denoting the correctly predicted exceedances as  $CP$ , the predicted exceedances as  $P$ , and the observed exceedances as  $O$ , we compute: the true predicted rate  $TPR = CP/O$ , the false positive rate  $FPR = (P - CP)/(N - O)$ , the false alarm rate  $FA = (P - CP)/P$ , and the success index  $SI = TPR - FPR$ .

$TPR$  and  $SI$  assume their best value at 1, while the best value for  $FPR$  and  $FA$  is 0.

To give robust estimates of performance indicators, a  $k$ -fold cross-validation is performed on available data, considering alternatively each year of data as the testing set and the remaining instances as the training set. Finally, results obtained on each testing set are averaged to obtain the performance indicators presented in the next section.

## 4. Results and discussion

Input variables selected for each of the four models (1 and 2 days ahead predictions at the two station sites) are shown in table 2, while table 3 shows the prediction performances.

Four main variables were selected in all input sets: the mean  $\text{PM}_{10}$  values of the last hours of day  $t$ , mean concentrations of  $\text{SO}_2$ , pressure and temperature computed on different time windows of day  $t$ . Concentration trends of pollutants on the same time windows resulted useful only for the one day forecast, while for the 2 days forecast horizon micrometeorological indicators and qualitative information about traffic fluxes seem to play a major role. Rainfall and wind speed measures, which are the main physical agents responsible of particulate dispersion, give only a slight enhancement in forecast quality, probably because the same information is contained in the previously selected variables. The potential benefits of accurate meteorological forecasts for days  $t+1$  and  $t+2$  as inputs for our models is still being investigated.

The true/predicted correlation for one-days-ahead predictions ranges from .82 to .86, while two-days-ahead predictors reach .71 on the same indicator.

With respect to the 1 day ahead forecast, the model for the Limito station performs better than the Milan-Juvara one, with a true/predicted correlation about 4% higher and a mean absolute error about 15% lower. Also, the success index on threshold exceedance is 5% higher for Limito. On the 2 days ahead forecast, such differences are limited to 10% in MAE and 2% in the success index.

Table 2: Selected inputs for each forecast model. The symbols  $\mu$  and  $?$  refer to the mean and difference operators and values within square brackets specify the interval (hours) of aggregation.

| Model                                | Selected inputs                                                                                                                                                                                  |
|--------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Milan-Juvara<br/>1 day ahead</b>  | $\mu(\text{PM}_{10}[18-24])$ , $\mu(\text{SO}_2[13-24])$ , $\mu(\text{Pressure}[1-24])$ , $\mu(\text{Temperature}[4-8])$ , $\mu(\text{Wind speed}[18-24])$ , $?(PM_{10}[18-24])$ , Day type(t+1) |
| <b>Milan-Juvara<br/>2 days ahead</b> | $\mu(\text{PM}_{10}[18-24])$ , $\mu(\text{SO}_2[8-24])$ , $\mu(\text{Pressure}[18-24])$ , $\mu(\text{Temperature}[4-8])$ , Height of mixing layer(t), Averaged flux of traffic(t+2)              |
| <b>Limito<br/>1 day ahead</b>        | $\mu(\text{PM}_{10}[18-24])$ , $\mu(\text{SO}_2[8-24])$ , $\mu(\text{Pressure}[1-24])$ , $\mu(\text{Temperature}[2-9])$ , (Wind speed[18-24]), Height of mixing layer(t), $?(PM_{10}[18-24])$    |
| <b>Limito<br/>2 days ahead</b>       | $\mu(\text{PM}_{10}[18-24])$ , $\mu(\text{SO}_2[1-24])$ , $\mu(\text{Pressure}[18-24])$ , $\mu(\text{Temperature}[4-8])$ , Height of mixing layer(t), Average flux of traffic(t+2)               |

Table 3: Prediction performances

|                                    | Milan-Juvara |              | Limito      |              |
|------------------------------------|--------------|--------------|-------------|--------------|
|                                    | 1 day ahead  | 2 days ahead | 1 day ahead | 2 days ahead |
| <b>Average goodness indicators</b> |              |              |             |              |
| <b>?</b>                           | 0.82         | 0.70         | 0.86        | 0.71         |
| <b>MAE</b>                         | 10.41        | 13.53        | 8.73        | 12.0         |
| <b>MBE</b>                         | 1.11         | 0.33         | 0.03        | -0.37        |
| <b>d</b>                           | 0.89         | 0.80         | 0.92        | 0.80         |
| <b>Threshold indicators</b>        |              |              |             |              |
| <b>TPR</b>                         | 0.75         | 0.70         | 0.79        | 0.68         |
| <b>FPR</b>                         | 0.10         | 0.14         | 0.08        | 0.10         |
| <b>FA</b>                          | 0.23         | 0.31         | 0.22        | 0.31         |
| <b>SI</b>                          | 0.66         | 0.56         | 0.71        | 0.58         |

The reliability of the forecast makes it potentially a valuable tool to support the warning system of the municipality, aimed at reducing the exposition, and thus the potential health damages, of sensible citizens (e.g. elderly, asthmatics). Furthermore, the dependence of the forecast from traffic fluxes opens the possibility of investigating policy decisions for pollution control.

## 5. References

- AMA, 2005. Rapporto sulla qualità dell'aria del Comune di Milano. Comune di Milano.
- Birattari, M., Bontempi, G., Bersini, H., 1999. Lazy learning for modelling and control design. *Int. J. of Control*, 72, 643-658.
- Corani, G. 2005. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling*, 185, 513-529.
- Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., Greig, A., Foxall, R., Eben, K., Chatterton, T., Vondracek, J., Richter, M., Dostal, M., Bertuccio, L., Kolehmainen, M., Doyle, M., 2003. A rigorous inter-comparison of ground-level ozone predictions. *Atmospheric Environment*, 37, 3237-3253.