# Bayesian models of the $PM_{10}$ atmospheric urban pollution

M. Cossentino[1], F.M. Raimondi[2] & M.C. Vitale[2]

[1] *Dipartimento di Ingegneria Elettrica, University of Palermo, Italy.*

[2] *Dipartimento di Ingegneria Automatica e Informatica, University of Palermo, Italy.*

## Abstract

In this paper we illustrate a forecast method of atmospheric pollution critical events caused by particulate matter (specifically $PM_{10}$) based upon the application of Bayesian networks.
These Bayesian networks model the temporal series of the pollutant during the day and the influence that meteorological parameters have upon them.
Each network has received some evidences (coming from historical records or meteorological forecasts) and used them to calculate its own forecast. Typical inputs of the networks have been the pollutant concentration at a certain hour and the meteorological parameters at the further hours of the day. The output provided by the networks is the estimate of the probability of reaching a certain pollutant level in the various hours of the day.
The results we have obtained in the prevision of the concentration of pollutant in the medium term are satisfactory and this approach can be profitably used to foresee critical episodes.

## 1 Introduction

Many different classical mathematical approaches to the problem of modelling atmospheric pollution are present in literature [1], [2], [3], [4]. Artificial intelligence and soft computing techniques has been recently introduced in similar problems obtaining good results. Starting from this observation we decided to evaluate the results that the bayesian networks could achieve in such a

context. Our goal was to forecast critical atmospheric pollution events, with particular regards to the particulate matter ($PM_{10}$), that reaches relevant concentrations in the city of Palermo.
We have decided to use the bayesian networks because atmospheric pollution is particularly dependent from meteorological factors that can be properly dealt with statistical models.
We have tested several different configurations in order to investigate the relevance of different factors such as the topology of the net, the number of historical data, the evidence provided and so on.
In the next paragraph we discuss the characteristics of the $PM_{10}$ pollutant; in the third paragraph we analyse the aspects that condition the emission and the diffusion of the $PM_{10}$ in the city of Palermo.
The fourth paragraph is a brief overview of the bayesian networks and of some related theoretical issues. In paragraph five we propose the different models that we have developed and then in the last paragraph, finally, the experimental results deriving from these model are analysed.

## 2 The $PM_{10}$ pollutant

The particulate matter is a pollutant that has drawn particularly the researcher's attention in the last years, and as a result of this interest various studies have been produced which document its dangerousness [5].
These pollutants are a mixture of solid and liquid particles that have different size (or dimension), composition and origin. This mixture is formed of chemical, organic, inorganic, mineral and vegetable compounds, that are natural or artificial, different both for their chemical nature and for physical behaviour.
Generally the particulate matter is classified according to the particles diameter (μm) and its concentration ($\mu g/m^3$). The particles diameter can change from 0,005 μm to 100 μm.
The Italian law regulates the presence of particulate $PM_{10}$ in the air. The $PM_{10}$ has a diameter that is lower than 10 μm, and it includes a thinner particulate's subgroup, that is called $PM_{2,5}$ (with a diameter lower than 2,5 μm).
The $PM_{10}$ can penetrate in the higher respiratory tract (thoracic cavity), whereas the $PM_{2,5}$ can arrive at the lungs. Moreover this kind of pollutant is the vehicle of irritant gasses, other toxic substances ($SO_2$, $NO_x$, carbon, lead, cadmium, arsenic, nickel, etc) and of cancerogenic substances (aromatic polycyclic hydrocarbons, nitro-compounds, aldehydes, etc.).
The natural sources of particulate matter are volcano eruptions, rocks erosion due to the wind, pollen or other vegetable materials dispersion. The particulate matter pollution produced by human activities, mostly derives from industrial plant, heating systems, and, especially in urban areas, from the wear of tyres and brakes and incomplete combustions related to the intense vehicular traffic.
The Italian law have established an annual threshold defined "goal of quality" that is the annual average concentration value. This value has been fixed at 40 $\mu g/m^3$.

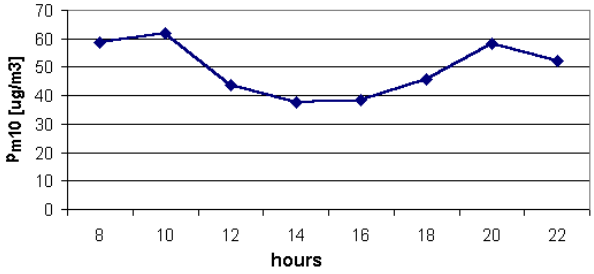


Figure 1: The average values of the $PM_{10}$ from 8 A.M. to 10 P.M. in the 1999

## 3 $PM_{10}$ Pollution in Palermo

The $PM_{10}$ concentration, in the city of Palermo, from 1997 to today, shows a significant variation of its average values. With the introduction of catalytic converter and some government financial incentive to the purchase of new cars, a considerable variation of fleet of cars and of vehicles typology has occurred. The modern cars use the so-called green petrol rather than the red petrol and therefore their pollutant emissions are greatly different.
Until 1997, in Palermo, a great quantity of CO (carbon monoxide) and lead have been diffused. After the year 1997, the concentration of these compounds is strongly decreased while other pollutants such as benzene, particulate matter (PM) and nitrogen oxide are much more present [6], [7].
For these reasons only data from the last three years could be taken into account for a study of the situation.
In Figure 1 the average values of the $PM_{10}$ from 8 A.M. to 10 P.M. in the year 1999 are shown. These values are almost unchanged in all years from 1998 to 2000 and therefore we decided to base our study only on the set of values of 1999.
Of this data we have only used records referring to the working days ignoring the festive days because the pollutant concentration has proved to be too much different not only in the level but also in the hourly distribution.
The meteorological parameters that we have considered are, obviously, those related with the diffusion of the $PM_{10}$. They turn out to be: the average speed of the wind, the direction of the wind, the amount of rains and the humidity.

## 4 Bayesian networks

The bayesian networks [8], [9], also called belief networks, fall within the ambit of classic probability theory that allows to deal with systems affected by uncertainty.
The basic concept of the bayesian approach regards the conditional probability that indicates the probability of the event A given the event B.

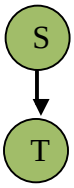



Figure 2: A simple bayesian network.

Table 1: the states associated with the variables of the network of Figure 2.

| Variable | States |
|---|---|
| S = solar radiation | Summer, spring (=autumn), winter |
| T = temperature | High, moderate, low |

The probability of the conjunct event is, instead, the probability that the events A and B are simultaneously true.
The bayesian networks allow us to represent the joint probability distribution for the environment of interest and their name descends from the Bayes theorem that is the fundamental mathematical tool in order to update the values of the networks when they are subjected to some new evidences:

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}. \qquad (1)$$

where P(A) is the probability of the event A, P(B) is the probability of the event B and P(B|A) is the conditional probability.
The bayesian approach is based upon the probability assigned to an event as a consequence of the current knowledge (*inference process*), we can say that a bayesian network provides a complete description of the application domain in the form of a conditional probability distribution.
A bayesian network is a directed acyclic graph in which each link is directed from a parent node to its child. Each node represents a variable of the domain and the links represent causal dependencies among the variables [10].
The possible values of each variable are divided into some intervals (of a given width) that are associated to a set of states (see Table 1). Evidence is called the a-priori information about the degree of certainty assigned to the possible states of a variable (see Table 2).
Each variable without parents has an a-priori probability table, while each variable A with some parents $B_1,\ldots,B_n$ has a conditional probability table that expresses the joint probability (i.e. P(T=high $\cap$ S=summer)=0.6) [11] (Table 3).

Table 2: The a-priori probabilities for the node S of the network in Figure 1

| **Solar Radiation** | | |
|---|---|---|
| Summer | Spring | Winter |
| 0.3 | 0.6 | 1 |

Table 3: Conditional probabilities for the node T of Figure 1

| | **Temperature** | | |
|---|---|---|---|
| *Solar Rad.* | High | Mod. | Low |
| Summer | 0.6 | 0.2 | 0.1 |
| Spring | 0.3 | 0.6 | 0.3 |
| Winter | 0.1 | 0.2 | 0.6 |

A generic element of the conditional probability table of Table 3 is the probability that occurs a combination of two specific values of the variables T and S; more in general, we can express it in the form: $P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n)$. Such an elements value is given by:

$$P(x_1,...,x_n)=\prod_{i=1}^{n} P(x_i|\text{Parents}(X_i)). \quad (2)$$

.

The construction of a bayesian network [9] can be performed through the following phases:

1) A set of variables $A_i$ (i=1, …, n) is chosen to describe the system;
2) To each variable $A_i$ is associated a node of the network;
3) The set Parents(Ai) of the nodes parents of $A_i$ is fixed;
4) The table of the conditioned probabilities for $A_i$ is defined often by the learning of a set of cases.

## 5 The proposed bayesian model

The model of networks that have been realized come down from the consideration that the concentration of pollutant at each hour derives from the concentration in the previous hours, the meteorological factors and the amount of emitted polluting.

It has not been possible in any way to estimate the amount of emitted pollutant because in Palermo the traffic volume that is considered the main responsible of pollution has never been studied in a systematic, quantitative manner.

Therefore we have considered only the following variables: concentration of the $PM_{10}$, speed of the wind, direction of the wind, humidity and amount of the precipitations.

We have decided to carry out a study of the evolution of these variables in intervals of two hours between 8 and 22 during the working days of the week.

Our aim is to achieve a forecast in the short-medium term and therefore we have decided to suppose to know the pollutant concentration at 8 A.M. and the forecast of the meteorological parameters for the rest of the day.

The networks we have tried are composed of *n* nodes: *m* are the nodes that represent the concentration of pollutant at the several hours, *k* are the nodes that represent the meteorological parameters.

A subset of *i* elements is assumed to be a set of evidences among the *k* meteorological nodes. These evidences are taken from the values of the test set and are provided to the network together with the corresponding pollutant concentration at 8 A.M. (in the following referred as PM10_8).

A simple network is represented in Figure 1 where 8 nodes of $PM_{10}$ and four nodes of meteorological parameters are present. These latter are parents of the

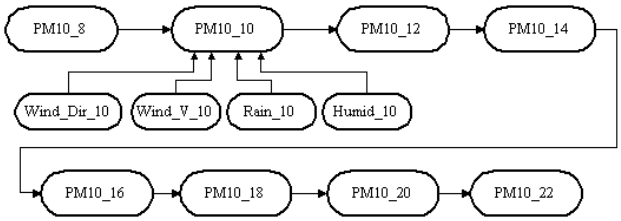



Figure 3: An example of network with meteorological parameters at 10 A.M.

node of $PM_{10}$ that is correspondent in time. Both meteorological nodes and the PM10_8 node are evidences of the network.

In choosing the topologies of the network and its parameters a precise process has been followed. It is composed of the following steps:

1) the topology of the network is chosen;
2) the size of the variables intervals are chosen and associated to the nodes. When a node represents a continuous quantity (like the concentration of a pollutant), it is necessary to divide its range of definition into a discrete set of values and then to associate the resulting intervals to the states of the node);
3) The inference process is performed (the new probabilities associated to the nodes of the network are calculated on the base of the historical data introduced in the network);
4) the set of evidences of the network is fixed (the number and the type of evidences, are key factors for the quality of the results and therefore each network topology has been tested with several different set of evidences).
5) the results are evaluated and compared with the previous ones (the results of each network are compared using diagrams and numerical quality indexes).

We have tested several network topologies and each of them with different configurations of the most important parameters (for example the width of the intervals of discretisation, the number of evidences, and so on).

The first network topology is a simple one (Figure 1). It is composed of 11 nodes, eight of which represent the $PM_{10}$ values between 8 A.M and 22 P.M; the other nodes represent the meteorological parameters like average speed of the wind (Wind_V), amount of the precipitations and direction of the wind (Wind_W) at 10 A.M..

We have supposed that these parameters are available by meteorological forecasts.

Some links connect the PM10_8, Wind_V, Rains, Wind_Dir nodes to the PM10_10 node. These connections give evidence of the influence that the meteorological factors at 10 A.M. and the concentration of pollutant at 8 A.M. have on the value of the $PM_{10}$ at 10. It is to be noted that a value of a node at 10 is

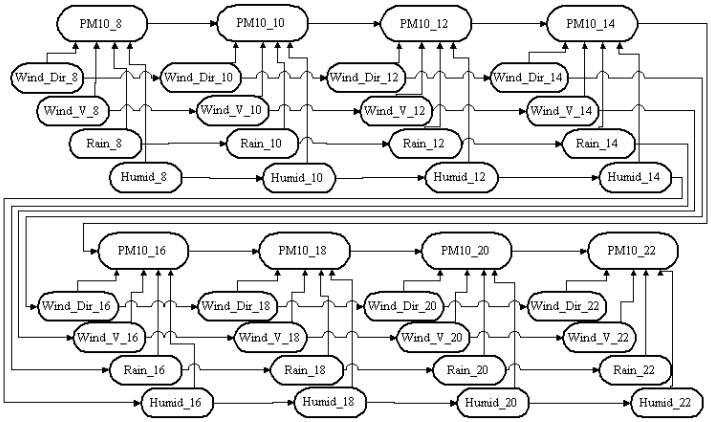



Figure 4: One of the most complete networks

the average of the values of the related variables in the previous two hours (from 8 to 10).

The links of the $PM_{10}$ nodes at different times take into account the obvious dependence of the values of the pollutant at a certain hour on the values of the variable in the two hours before.

In Figure 2 we can see one of the most complex networks. It includes the concentration of the $PM_{10}$, the speed of the wind, the direction of the wind, the amount of the rains and relative humidity between 8 and 22.

The intermediate configurations allowed to study the influence of the different meteorological parameters on the quality of the resulting forecast.

For example we have noticed that the rain nodes give a contribution that is often negligible. This is probably due to the fact that days with consisting rain are quite rare in Palermo and therefore they constitute a insignificant occurrence from the statistical point of view.

On the contrary, humidity has proved to be a very important parameter; it often reaches high values and considerably conditions the pollutant fall.

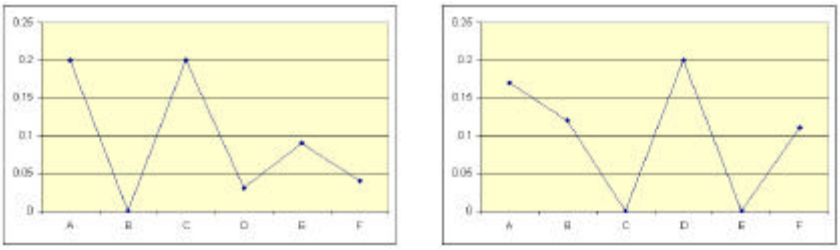
## 6 Results

Analysing the behaviour of the networks we have deduced the role that some factors have on the goodness of the results.

For example networks with small width of the intervals associated with different states of the nodes have given, in general, good results but they have shown difficulties with great and rapid changes of pollutant concentration.

This behaviour can be observed in the errors of the same network obtained with different values of the discretisation ranges (Figure 5).

The second network has a smaller range and it gives better results.




(a) (b)

Figure 5: Daily relative error for the network of figure 3 in some days of the test set. The discretisation range of the $PM_{10}$ nodes is 30 $\mu g/m^3$ (a) and 10 $\mu g/m^3$ (b)

We have also noticed that networks with a greater knowledge of meteorological parameters, often show unstable results. Probably this is due to a lack in the dimension of the training set.

Finally, we could drawn the following considerations:

1) The intervals width of the $PM_{10}$ nodes is critical for the behaviour of the whole network. Wider intervals give place to a good performance in correspondence of some rapid variations of the pollutant, while smaller intervals cause a better precision in following small variations.
2) The knowledge that the network has of the system is, obviously, important. A network with few nodes is expected to give worse results than a network with a greater number of nodes because the latter has a better description of the system and has more available inputs to count on. Unfortunately, in many cases it is not so probably, because the dimension of the required training set increase rapidly with the number of the network nodes.
3) Another factor that has a relevant influence in the quality of the results is the number of the evidences that are supplied to the network. This theoretical issue, has not been confirmed by the experimental results. Sometimes, networks with less evidences have given better results than the homologous network with more evidences.

The network that has given the best results is shown in Figure 3. Its $PM_{10}$ states have a width of 10$\mu g/m^3$ and the provided evidences are: speed and direction of the wind, humidity and amount of rains at 10 A.M., concentration of $PM_{10}$ at 8 A.M.

In Figure 5 the results obtained in some days of the test set are reported.

From these experiments, we have deduced that the bayesian networks can provide good results in the forecast of atmospheric pollutants concentrations in the medium term.

Particularly we can notice that network topologies with a few parameters (i.e.: speed and direction of the wind, humidity and rains at 10 A.M., concentration of $PM_{10}$ at 8 A.M.) can have a valid application in the forecast of a pollutant trend in the 14 successive hours.

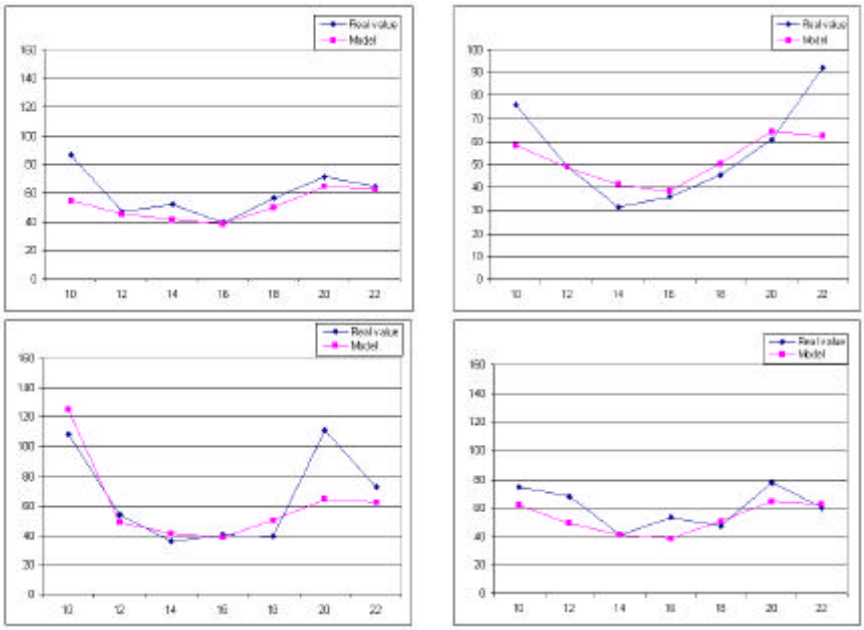


Figure 6: The results of the network of Figure 3 for some days of the test set

It is possible in fact, to use an automatic system that uses the value of $PM_{10}$ at 8 and the forecasts for the two successive hours of the meteorological parameters to provide an estimated trend of the $PM_{10}$ until 22.
For this reason we have extensively tested networks with such a configuration of evidences and, with a network like that of Figure 3, we have obtained the results shown of Figure 6.

## 7 Conclusions

A method for the forecast of atmospheric pollution critical events has been developed using the bayesian networks.
The typical input of a network is the pollutant concentration at a given hour of the day and the meteorological parameters to the successive hours.
The forecast consists of the probability that the $PM_{10}$ reaches a certain pollution threshold.
A great number of tests, based on data recorded by several monitoring stations in the city of Palermo, have been performed. From their results, it is possible to deduce that the behavior of the networks is conditioned by the width of the intervals associated to the states of various nodes, by the number of these nodes, by the number of the historical series available and by the number of the evidences supplied to the network. An optimal network topology has been identified and the results obtained, in the medium term, are particularly interesting.

The best results have been provided by a configuration whose evidences are the parents of the node PM10_10 ($PM_{10}$ at 10 A.M.). These parents are the values of $PM_{10}$ at 8.00 and meteorological parameters at 10.00.

## References


[1] Zannetti, P., *Air pollution modeling*, Computational Mechanics Publications: Southhampton, Boston, 1990.

[2] Finzi, G. & Brusasca, G., *La qualità dell'aria. Modelli previsionali e gestionali*, Masson: Milano, 1991.

[3] Raimondi, F.M., Rando, F., Vitale, M.C. & Calcara A.M.V., Short-time fuzzy DAP predictor for air pollution due to vehicular traffic. *Proc. of the 1st Int. Conf. on measurements and modelling in environmental pollution (MMEP 97)*, Computational Mechanics Publications: Southampton and Boston, 1997.

[4] Raimondi, F.M., Rando, F., Vitale, M.C. & Calcara A.M.V., A Short-term air pollution predictor for urban areas with complex orography. Application to the town of Palermo. *Proc. of the 1st Int. Conf. on measurements and modelling in environmental pollution (MMEP 97)*, Computational Mechanics Publications: Southampton and Boston, 1997.

[5] Mazzali, P., "*Inquinamento atmosferico: Origine, Prevenzione, Controllo*", Pitagora Editrice: Bologna, 1990.

[6] "*Il rilevamento dell'inquinamento acustico ed atmosferico nel comune di Palermo. 1a Relazione*, AMIA (Azienda Municipalizzata Igiene Ambientale): Palermo, 1998.

[7] "*Il rilevamento dell'inquinamento acustico ed atmosferico nel comune di Palermo. 2a Relazione*, AMIA (Azienda Municipalizzata Igiene Ambientale): Palermo, 1999.

[8] Pearl, J., Belief Networks revisited. *Artificial Intelligence*, n. 59, 1993.

[9] Russel, S. & Norvig, P., "*Artificial Intelligence. A modern approach*", Prentice Hall International Editions, 1995;

[10] Jensen, F.V., An introduction to bayesian networks, *Proc. of the 10th Conf. on Uncertainty in AI*, 1994.

[11] Drudzel, M.J., Some properties of Joint Probability Distributions. *Proc. Of the 10th Annual Conference on Uncertainty in Artificial Intelligence*, UAI-94: Seattle, Washington, pp. 187-194, 1994.