Multivariate Stochastic Models of Sulphur Dioxide Pollution in an Urban Area

G. Finzi, G. Fronza and A. Spirito
Istituto di Elettrotecnica ed Elettronica
Centro Teoria dei Sistemi C.N.R.
Politecnico, Milan, Italy

Three multivariate stochastic mathematical models of daily SO_2 pollution in an urban area (Milan, Italy) during the heating season (mid-October/end of March) are illustrated in the paper. Each model is characterized by a different number of external inputs. Precisely, the first model has no inputs (it is simply an autoregressive relationship), the second one has a temperature input (roughly accounting for emission), the third one has two inputs (temperature and wind speed). From each model a real-time predictor is derived, namely a recursive relationship which, at the end of each day, allows future pollution levels to be forecast on the basis of current concentration and meteorological measurements. The quality of the forecast is rather satisfactory, even in episode situations. The improvements in forecast performance when turning from a predictor with less external inputs to a predictor with more external inputs (i.e., when exploiting more information about meteorology) are also pointed out in the paper.

As an alternative to deterministic mathematical representations (mainly of the gaussian-type or based on the K-theory), stochastic mathematical models such as ARIMA (Auto Regressive Integrated Moving Average) or seasonal ARIMA have been largely used in order to fit time series of pollutant concentrations.[1-4]

In accordance with the techniques recommended by Box and Jenkins,[5] from an ARIMA model a real-time predictor can be derived, namely a recursive relationship which, at the beginning of each time interval, supplies the "optimal" forecast of future concentrations on the basis of concentration data recorded during the previous intervals.

However, the absence of physical inputs such as meteorological and emission variables makes ARIMA predictors unable to give a satisfactory forecast performance in presence of pollution episodes. In fact, such rapid increases and subsequent decreases of pollutant concentrations can be explained only by putting into evidence the existence of a particular meteorological and/or emission situation. Therefore, a better real-time predictor can be derived by the so-called ARIMAX (ARIMA with eXogeneous inputs) stochastic mathematical representations.[5,6] In such models, the average concentration in a certain interval is expressed as a linear combination of past average concentrations plus linear combinations of present and past physical inputs plus the noise term.[7-9]

Because of their simple structure and satisfactory performance, ARIMAX represent a serious alternative to episode description by means of the advection-diffusion equation, which must usually be integrated through a complex numerical scheme and requires extremely detailed information on meteorology and emission. Actually, a comparison between the two approaches is improper,[10] because of the different model objectives. The advection-diffusion equation aims at representing the time behavior of the entire concentration field, while existing ARIMAX (as well as ARIMA) applications supply only a local or aggregate description of the pollution phenomenon. Precisely, they are univariate models, namely they aim at representing the dynamics of a single variable, either the concentration in a certain monitoring station or a spatial concentration average (the DAP or Dosage Area Product[11]).

An intermediate detail of description of the phenomenon of pollutant dispersion is given by the stochastic models illustrated in the present paper, together with their application to a real case (sulphur dioxide pollution in the heating season in Milan, Italy). Specifically, they are multivariate stochastic models which describe the dynamics of a vector of three variables, taken as representative of pollution (the vector of daily DAPs in the three sectors of the area under consideration).

A final remark concerns the possibility of using real-time predictions for air quality control purposes. Case studies and real world applications of emission control based on forecasts (although not supplied by a mathematical predictor) have already been illustrated in the literature.[12,13] Such applications of intermittent control policies (emissions are reduced only when *an episode* is forecast) mainly concern pollution due to a few sources, although interesting implementations to urban air pollution cases can be found in Japanese cities like Osaka.[14] A control action of this kind is not presently feasible in Milan, so the pollution predictor described in this paper will only be used by the public authority for alarm purposes.

Reprinted from **APCA JOURNAL**, *Vol. 30, No. 11, November 1980*

Description of the Area, Data Set and DAP Vector

The circular region R under consideration is shown in Figure 1a, together with its subdivision into three sectors R_1, R_2, R_3 and the network of ten SO_2 monitoring stations, each supplying hourly averages.

The city is situated in a completely flat area, so there is no orographic effect.

Hourly temperature, wind speed, and direction are measured by one meteorological station in the center of region R. Particularly in the heating season, Milan is affected by a very weak circulation.[15] Moreover in such season there are often many days of calm and inversion, which lead to high accumulation of pollutant (hourly averages sometimes over 1 ppm).

The frequencies of wind direction in the time period analyzed in this study are shown by the wind rose of Figure 1b (however, heating periods of different years are characterized by the very same behaviour).

As for emission, in winter sulphur dioxide in R is practically due only to residential heating, since there are neither significant industrial sources in R nor significant transport from external industrial zones. In particular the industrialized area of Sesto is located in the northeast, the direction which corresponds to the most infrequent wind. Day–by–day emission rates are obviously not measured, so emission is only indirectly taken into account in the models.

In order to obtain the time series, of the variables taken as pollution representative, the daily dosage data recorded by the monitoring stations have been manipulated as follows.

Consider a coordinate system with the origin in the center of region R, the X-axis and the Y-axis directed towards the east and the north respectively. Furthermore let (x_i,y_i), i = 1,2, . . .10, be the coordinate of the i-th station and let $D_i(k)$ denote the k-th day dosage in the i-th station.

The Dosage Area Product, $DAP_j(k)$, over the j-th sector R_j (j = 1,2,3) of the region R is the integral of the dosage over R_j. An approximate evaluation of such integral by the measured variables $D_i(k)$ can be obtained by the following method.

i. Subdivide R by a thick grid.

ii. Assume that in each square of the grid the dosage is the same as in the station nearest to the square (i.e. giving each station a polygonal "area of influence"). There turns out an approximation of the type

$$DAP_j(k) = \frac{1}{A_j} \sum_{i=1}^{10} a_i^j D_i(k) \qquad (1)$$

where

A_j = area of R_j

a_i^j = area of the surface common to sector R_j and to the polygon covering the grid squares near station i.

By considering the daily dosage data recorded in the period October 16, 1975–March 31, 1976–October 16, 1976–March 31, 1977 by the ten monitoring stations and by applying Eq. (1) to such data, a time series of the vector of interest $DAP\ (k) = |DAP_1(k), DAP_2(k), DAP_3(k)|'$ ($'$ = vector transposition symbol) has been obtained. Part of such time series (Oct. 16, 1975–Mar. 31, 1976) from now onwards called "historical DAP series," together with the series recorded by the meteorological station in the same period, has made it possible to estimate the parameters of the model of the DAP vector described in the next sections. The remainder of the series (Oct. 16, 1976–Mar. 31, 1977) has been used for testing the forecast performance of the predictors derived from the models.

An AR(1) Model of the DAP Vector (No Physical Inputs)

First the "historical DAP time series" mentioned above has been assumed to be a (finite) realization of a three-variate stationary random process $\{DAP(k)\}_k$ (in fact there is no evidence of cycles or trends in the series).

Then the DAP process has been represented by the following three-variate ARMA (Auto-Regressive Moving Average = stationary ARIMA) model:

$$DAP(k+1) = \sum_{j=1}^{p} \phi_j DAP(k-j+1) + \epsilon(k+1) + \sum_{m=1}^{q} \psi_m \epsilon(k-m+1) \qquad (2)$$

where

ϕ_j, ψ_m = 3×3 – matrices of model parameters;

p,q = model orders;

and $\{\epsilon(k)\}_k$ is a three-variate, purely random, zero mean pro-

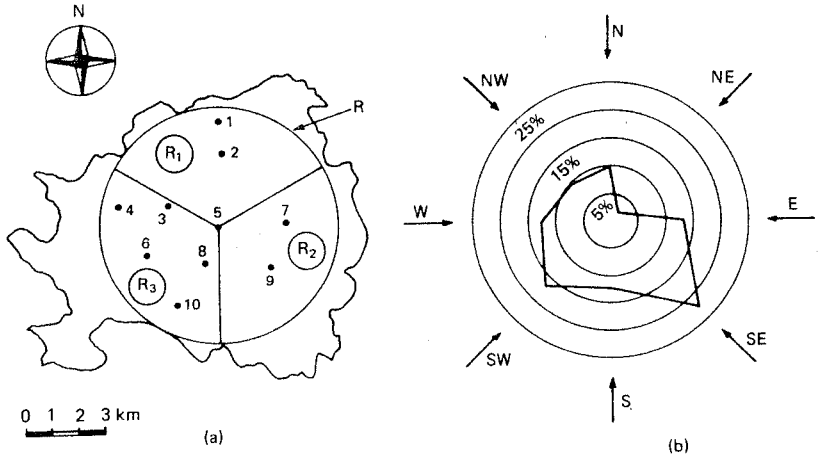

Figure 1 a. The region under consideration, the three sectors and the SO_2 monitoring network. **b.** Wind rose in the heating season.

Table I. a. *DAP* means and standard deviations. b. "Episode process" means and standard deviations.

J	1	2	3
a)			
μ_j [ppm]	0.208	0.183	0.168
σ_j [ppm]	0.115	0.101	0.089
b)			
μ_j^E [ppm]	0.420	0.357	0.340
σ_j^E [ppm]	0.105	0.087	0.081

cess (white noise), namely a gaussian random process characterized by the following statistics ($E[\cdot]$ = expectation operator)

$$E[\epsilon(k)] = 0$$

$$E[\epsilon(k)\,\epsilon'(k+\tau)] = \begin{cases} 0 & \tau \neq 0 \\ \neq 0 & \tau = 0 \end{cases}$$

The most suitable in the class (2) has turned out to be the AR(1) (Auto Regressive of order 1), namely

$$DAP(k+1) = \phi DAP(k) + \epsilon(k+1) \qquad (3)$$

In particular, the estimation of ϕ from the historical DAP series has been made through a standard fitting technique (Yule-Walker equations[5,16]).

In ARMA modelling it is also necessary to validate *a posteriori* ("diagnostic checking"[5]) the initial assumption concerning the form of the stochastic part of the model. This has been done for model (3) and actually $\{\epsilon(k)\}_k$ has turned out to be zero mean and uncorrelated (noise correlation tests have been performed through the analysis of the cumulative periodogram of the residual[5]).

The one day ahead predictor derived from model (3) is simply[5]

$$\hat{DAP}(k+1/k) = \phi DAP(k) \qquad (4)$$

where
$\hat{DAP}(k+1/k)$ = forecast of DAP $(k+1)$ made at the end of the k-th day.

Component by component, the forecast performance of the DAP predictor (4) has been measured by the following indexes:

θ_j = correlation between predicted and observed DAP_J;
θ_j^E = correlation between predicted and observed DAP_J during episodes;
σ_j'/μ_j = ratio between the standard deviation of the forecast error $\epsilon_j(k) = DAP_j(k) - \hat{DAP}_j(k)$ and the mean of $\{DAP_j(k)\}_k$;
η_j^E/μ_j^E = ratio between the mean square error of the forecast and the mean of $\{DAP_j(k)\}_k$ during episodes.

Specifically an episode has been defined as a time interval characterized by $DAP_j(k) > \mu_j + \sigma_j$ (means and standard deviations of DAP and of episodes process are reported in Table I).

The forecast performance in correspondence with the heating season 1976/77 is summarized in Table II (top part). It is only slightly better than the performance of the persistent DAP predictor ("DAP tomorrow will be the same as today") reported in the same Table (bottom part).

A Single Input ARX Model (Input:Temperature)

As stated earlier, pollution episodes can be explained only by models which put into evidence the role of emission and/or meteorology. Hence, a first refinement of model (3) has been made by introducing an external input representing overall emission from the sources in the area under consideration.

Of course, day-by-day emission from each residential heating source is not recorded. Hence, it has been assumed that overall emission in the $(k+1)$-th day can be somehow taken into account by introducing the average temperature $T(k)$ of the k-th day as model input. Precisely, the following ARX representation (Auto Regressive with eXogenous input) of the DAP vector has been considered:

$$DAP(k+1) = \phi^I DAP(k) + f(T(k)) + w(k+1) \qquad (5)$$

where ϕ^I is a matrix of model parameters, $\{w(k)\}_k$ is zero mean white noise and $f(T(k)) = |f_1(T(k))f_2(T(k))f_3(T(k))|'$, with

$$f_j(T(k)) = a_j/(b_j + T(k)) \qquad j = 1,2,3 \qquad (6)$$

Table II. Forecast performance of predictors with nonexogenous input.

Predictors	J	θ_j	θ_j^E	σ_J'/μ_J	η_J^E/μ_J^E
AR(1)	1	0.71	0.54	0.40	0.36
	2	0.75	0.65	0.38	0.33
	3	0.71	0.36	0.38	0.36
Persistent	1	0.68	0.53	0.44	0.36
	2	0.73	0.65	0.40	0.31
	3	0.69	0.33	0.41	0.37

The first remark about model (5) concerns the use of $T(k)$ instead of $T(k+1)$ as representative of emission in the $(k+1)$-th day. This is equivalent to postulate a one day delay in the reaction of people to temperature changes.

As for the form of the temperature input, since $f_j(T(k))$ is representative of emission in model (5), it must *a priori* be a decreasing function. Naturally, the choice of the hyperbolic form (6) is arbitrary, but it has proved satisfactory when running the predictor derived from Eq. (5).

The estimation of the parameters of model (6) (namely of the matrices ϕ^I, $a = |a_1\,a_2\,a_3|'$, $b = |b_1\,b_2\,b_3|'$) from historical DAP and temperature data 1975/76 gives some extra problem with respect to the estimation of ϕ in model (3). In fact, model (5) is nonlinear with respect to the b parameters (because of the form (6) assumed for $f_j(T(k))$). Hence, standard least squares fitting techniques[17] cannot be applied. So, a simple iterative parameter estimation procedure has been set up, based on the consideration that model (5) is linear with respect to ϕ^I and a. The general step of such procedure consists of the following operations:

i. a value of b is given by the previous step; in correspondence with such given value of b, optimal estimates of ϕ^I and a are found through standard least squares fitting;
ii. in correspondence with the values of ϕ^I and a supplied by i), b is changed in accordance with a gradient technique[18] in order to improve fitting.

Operations i. and ii. are iterated until the estimates of ϕ^I, a and b converge.

The real-time predictor derived from the ARX model (6) is:

$$\hat{DAP}\,(k+1|k) = \phi^I DAP(k) + f(T(k)) \qquad (7)$$

and its performance is shown in Table III (top part). The comparison with Table II points out the improvement with respect to predictor (4), which has no external inputs.

A Two-Input ARX Model (Inputs: Temperature and Wind Speed)

In order to obtain a further improvement of the forecast performance, average daily wind speed has been introduced into model (5), thus turning it into the following two-input ARX representation:

$$DAP(k+1) = \phi^{II} DAP(k) + f^{I}(T(k)) + cV(k+1) + p + \nu(k+1) \quad (8)$$

In model (8), ϕ^{II} is a matrix of parameters, $f^{I}(T(k))$ is the "emission representative" input, $c = |c_1 c_2 c_3|'$ and $p = |p_1 p_2 p_3|'$ are new vectors of parameters and $\{\nu(k)\}_k$ is white noise. Of course, parameter estimation for model (8) from historical *DAP*, temperature and wind speed data of the heating season 1975/76 has supplied a negative value for each component of c. The estimation has been carried out via a procedure quite similar to i., ii. illustrated in the previous section.

The predictor derived from model (8) is

$$\hat{DAP}(k+1|k) = \phi^{II} DAP(k) + f^{I}(T(k)) + cV(k+1) + p \quad (9)$$

namely the forecast of $DAP\ (k+1)$ depends upon $V(k+1)$ which clearly is not an available datum at the end of the k-th day, when the forecast is made. To overcome such difficulty, it would be necessary to set up a separate model and predictor of wind speed and subsequently introduce day by day wind speed forecasts as inputs for the DAP predictor (9).

Otherwise, one can consider the following two extreme situations:

a. *Trivial wind speed prediction.* It corresponds to assume persistent wind speed ($V(k+1) = V(k)$), i.e. to introduce $V(k)$ into the *DAP* predictor (9). This approach corresponds to the roughest treatment of the wind input into Eq. (9) and hence yields a lower bound for the forecast performance of the *DAP* predictor.

b. *Ideal wind speed predictor.* It corresponds to introduce the true $V(k+1)$ into Eq. (9), namely to forecast the future wind speed exactly. Of course, this procedure yields the upper bound for the forecast performance of the *DAP* predictor (9).

Upper and lower bounds in the Milan case are reported in Table III (bottom and middle part, respectively). Note that even the lower bound is an improvement with respect to the figures supplied previously.

Table III. Forecast performance of predictors with meteorological inputs.

Meteorological inputs to ARX predictors	J	θ_j	θ_j^E	σ_J/μ_J	η_j^E/μ_j^E
$T(k)$	1	0.76	0.61	0.36	0.34
	2	0.79	0.66	0.35	0.31
	3	0.75	0.42	0.36	0.32
$T(k), v(k)$	1	0.77	0.62	0.36	0.33
	2	0.80	0.67	0.34	0.29
	3	0.77	0.44	0.35	0.31
$T(k), v(k+1)$	1	0.84	0.69	0.32	0.28
	2	0.84	0.72	0.31	0.27
	3	0.82	0.50	0.31	0.28

Acknowledgment

This research has been supported by C. N. R.—Programma Finalizzato "Promozione della Qualità dell'Ambiente."

References

1. P. H. Merz, L. Y. Painter, and P. R. Ryason, "Aerometric data analysis. Time series analysis and forecast and an atmospheric smog diagram," *Atmos. Environ.* **6:** 319 (1972).
2. D. P. Chock, T. R. Terrel, and S. B. Levitt, "Time-series analysis of Riverside, California air quality data," *Atmos. Environ.* **9:** 978 (1975).
3. G. M. McCollister and K. R. Wilson, "Linear stochastic models for forecasting daily maxima and hourly concentrations of air pollutants," *Atmos. Environ.* **9:** 417 (1975).
4. G. C. Tiao, G. E. P. Box, and W. J. Hamming, "A statistical analysis of the Los Angeles ambient carbon monoxide data 1958–1972," *JAPCA* **25:** 1129 (1975).
5. G. E. P. Box and G. M. Jenkins, *Time Series Analysis, Forecasting and Control,* Holden-Day, San Francisco, 1970.
6. P. Young, P. Whitehead, "A recursive approach to time series analysis for multivariable systems," *Int. J. Control* **25:** 457 (1977).
7. G. Finzi, G. Fronza, and S. Rinaldi, "Stochastic modelling and forecast of the dosage area product," *Atmos. Environ.* **12:** 831 (1978).
8. G. Finzi, P. Zannetti, G. Fronza, and S. Rinaldi: "Real time prediction of SO_2 concentration in the Venetian Lagoon area," *Atmos. Environ.* **12:** 1249 (1979).
9. G. Finzi, G. Fronza, S. Rinaldi, and A. Spirito, "Prediction and Real-Time Control of SO_2 Pollution from a Power Plant," Air Pollution Control Association Annual Meeting, Houston, 1978.
10. G. Fronza, A. Spirito, and A. Tonielli, "Real-time Forecast of Air Pollution Episodes in the Venetian Region. Part 2, The Kalman Predictor," *Appl. Math. Modelling* (in press).
11. S. Duckworth and E. Kupchanko, "Air analysis: the standard dosage area product," *JAPCA* **17:** 379 (1967).
12. D. S. Shepard, "A load shifting model of air pollution control in the electric power industry," *JAPCA* **20:** 756 (1970).
13. J. M. Leavitt, S. B. Carpenter, J. P. Blackwell, and T. L. Montgomery, "Meteorological program for limiting power plant stack emissions," *JAPCA* **21:** 400 (1971).
14. T. Soeda, S. Omatu, "Real-time Control of Emissions in Japanese cities," Proceedings of the Workshop on Mathematical Models for Planning and Control of Air Quality (G. Fronza, P. Melli eds.) Laxenburg—Austria, 1979 (in press).
15. L. Santomauro, R. Gualdi, G. Tebaldi, "Modello di Simulazione della Diffusione Atmosferica Locale (Area Urbana di Milano)" CNR-AQ/3/1. Coll. del Prog. Fin. "Promozione della Qualità dell'Ambiente," 1978.
16. N. C. Matalas, "Mathematical assessment of synthetic hydrology," *Water Resource Res.* **3:** 937 (1967).
17. K. Y. Åstrom, P. Eykhoff, "System identification: a survey," *Automatica* **7:** 123 (1971).
18. D. G. Luenberger, *An Introduction to Linear and Nonlinear Programming,* Addison-Wesley, Reading, MA, 1973.