blished by PENTECH PRESS LIMITED
ndon : Plymouth

duced Photolitho by
ood Burn Limited
ridge & Esher

1977 APPLIED NUMERICAL MODELLING

Preprints
of the
International Conference on

APPLIED NUMERICAL MODELLING

11 – 15 July, 1977

UNIVERSITY OF SOUTHAMPTON,
SOUTHAMPTON SO9 5NH,
ENGLAND

Presented at the Int. Conference on "Applied Numerical Modelling", University of Southampton, 11-15 July, 1977 - 1

P - 10

II - 53

MODELLING AND FORECASTING SO_2 AIR POLLUTION LEVELS: A STATISTICAL APPROACH

P. Zannetti

IBM Scientific Center, Venice, Italy

ABSTRACT

This paper first reviews the previous work dealing with the subject of the statistical approach to the study of meteorological and air quality measurements. Different applications of such methodologies have suggested a certain number of statistical analyses which were applied to SO_2 and meteorological data recorded in the Venetian area. An investigation of the frequency distribution of the data was carried out together with the application of spectral analysis in order to understand the periodicity of the meteorological and air quality data. The main objective of this paper is the building of various stochastic models which take into account the daily cycle of SO_2 concentrations. Regression analysis is also applied in order to increase the performances of the SO_2 predictors by the use of meteorological input. The same methodologies are applied to a global air pollution index which can describe the average phenomenon in the entire Venetian area.

INTRODUCTION

By considering pollutant concentration at a particular monitoring point as a stochastic process, we gain the initial advantage of neglecting almost all of the problems regarding pollutant emissions and atmospheric diffusion. In this way, the concentration time series to be analysed may be thought as one particular realization of a stochastic process which is defined as a statistical phenomenon that evolves in time according to probabilistic laws (Box and Jenkins, 1970). This approach is a limited one, but shows its efficiency in the analysis and forecasting of time series. According to previous works on this subject, the statistical study of air pollution data has developed along four main directions: 1) analysis of frequency distributions of air quality data; 2) spectral analysis of meteorological and air quality time series; 3) multiple

regression; and 4) Box-Jenkins forecasting methodology.
As regards the first methodology, Larsen's works (e.g. Larsen, 1969) have identified two general laws which can be expressed in the following form: 1) concentrations are approximately lognormally distributed for all pollutants in all cities for all averaging times; and 2) the median concentration (50 percentile) is proportional to averaging time to an exponent. Lognormality of air quality data has been confirmed by a great number of experimental studies, and it has been possible to present a heuristic justification of this distribution(Kahn, 1973).
The second approach, spectral analysis, has been applied to the analysis of measured time series in order to both identify the main periodicities and to understand the correlations between air quality and meteorological pattern. Two interesting studies in this field (Tilley and McBean, 1973; Trivikrama et al., 1976) show the existence of the following main oscillations of SO_2 and wind speed data: semi-diurnal, diurnal and 3-3.5 day periods. Diurnal and semi-diurnal cycles are ascribed to local phenomena like the sea breeze, while the longest period oscillation is due to synoptic weather variations which are known to have a period close to 3.5 days in their study area (North-East America).
Using the third approach, there are many different measured variables which can be used as predictors to forecast pollution values for short term control: pollutant concentrations and local and more distant meteorological variables. By applying multiple regression it is possible to find the linear combination of these predictor variables which best forecasts pollution levels. A great number of multiple regression studies deal with the forecasting of oxidant levels by determination of the actual form of the oxidant-meteorological and/or oxidant-precursor dependence (Chock et al., 1974; Tiao et al., 1975).
As regards the last methodology it must be pointed out that by following the Box-Jenkins technique it is possible to use a methodology which shows its adaptive ability both in forecasting and in the handling highly correlated series. This method can take into account autoregressive (AR) and/or moving-average (MA) behaviour of a time series, thus reducing to a minimum the number of predictor parameters to be estimated. This approach has been utilized in many air pollution analyses and forecasting studies (Mertz et al., 1972; McCollister and Wilson, 1975; Chock et al., 1975).
This paper describes the area under investigation (Venetian area), the data collected, the current results obtained by applying deterministic diffusion models in order to simulate air pollution dynamics, and then describes the application of statistical models to the Venice data for analysis and forecasting problems. A particular emphasis is placed on the definition and use of stochastic predictors in order to supply industries with a good control tool for the reduction of their emissions as a consequence of high forecasted values of pollution levels. These predictors are applied both in fitting and in forecasting SO_2 values. In some of the models different meteorological inputs have been taken into account by multiple linear regression with

the SO_2 data.

VENETIAN AREA

Geography

The area of investigation (Figure 1) is a section of the Venetian Lagoon located in the northeastern part of Italy at the upper shore of the Adriatic Sea, from which the Lagoon is separated by two narrow strips of land: Lido and Pellestrina. It includes the urban centers of Mestre, Marghera and Venice, and the heavily industrialized area of Porto Marghera. The urban centers of Mestre and Marghera are situated on the mainland and have a surface area of about 10 km^2. Close by is a large industrial area of about 20 km^2, whose main activities include oil-refining, petrochemicals production, metallurgical processing of iron and other metals, and production of electric energy. Five km from the mainland, in the middle of the Lagoon, is the historical center of Venice, covering an area of 6 km^2 and standing on a cluster of small islans separated by a network of small canals and interconnected by many bridges. The region is on the extreme end of the Padana plain and is essentially flat.

Meteorology

Apart from few topographical effects, the complexity of the local meteorology (Runca and Zannetti, 1973; Zannetti et al.,

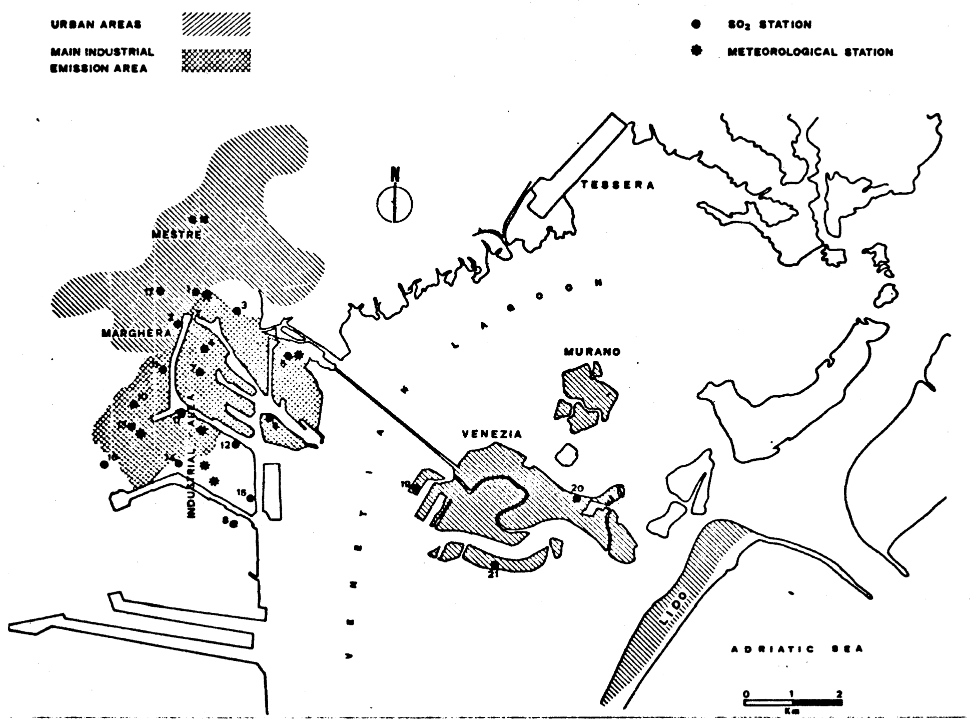

Figure 1 The Venetian Lagoon and the stations of the Ente Zona monitoring network.

in press) is mainly due to the presence of different types of surfaces. Specifically, the Venetian region has a Mediterranean climate with the peculiarity of a high average relative humidity (about 80%). The prevailing wind direction is from the NE, however there are important seasonal variations. In particular, the frequency of the wind blowing from the industrial area towards Venice (i.e., from the W and the NW) is generally low although higher in winter. Low speed (≤ 3 m/s) conditions predominate in the area. A sea breeze effect has been detected by examining the evolution of the hourly wind roses corresponding to periods characterized by speeds lower than 3 m/s. Atmospheric stability is neutral on the average, however in winter the frequency of stable conditions increases, mainly in correspondence with winds blowing from the mainland towards Venice.

Data

Two networks of monitoring stations, measuring both air quality and meteorological data, have been set up in the last four years. However, the data used in the present analysis have been provided only by the network installed in September 1974 by the local Industries Committee (Ente Zona Industriale di Porto Marghera). This network consists of 21 stations (Figure 1) measuring ground level SO_2 and 4 meteorological stations in the industrial area. The data are supplied in the form of thirty-minute average at each station. The meteorological stations supply wind speed and direction, temperature, pressure, relative umidity, rainfall and stability according to Pasquill's classification. For the period 1973-74 the total SO_2 emission due to the industries has been estimated about 160,000 tons per year, while the urban emission, concerning 5-6 winter months each year, has been evaluated about 10,000 tons per year. As a matter of fact, SO_2 averages exibit a substantial increase during winter due to both the contribution of domestic emissions and to the worse meteorological conditions.

Previous studies

The first analysis of the air pollution problem in the Venetian area has followed the deterministic (or mechanistic) approach. Precisely, a climatological gaussian model has been set up in order to describe seasonal and yearly SO_2 averages (Runca et al., 1976-B). This model has given a satisfactory fitting of the data and can be used both in plant location problems involving pollution, and for long-term forecast. A satisfactory simulation of daily SO_2 averages has been also provided by application of a three-dimensional code based on numerical integration of diffusion-convection equation (Runca et al., 1976-A). Recently, the problem of short-term prediction oriented to a real-time control or, at least, to an alert system, has suggested to follow statistical techniques in order to provide stochastic SO_2 predictors (Zannetti, in press).

STATISTICAL MODELS APPLICATION TO THE VENICE AREA

SO_2 frequency distribution

According to the analysis carried out for a large number of cases (Hunt, 1972; Bencala and Seinfeld, 1976), the lognormal distribution yields a satisfactory fit of the data of different pollutants in many sites and in correspondence with all averaging times. In the Venetian case, for each season the lognormality assumption has been first tested and generally accepted for the hourly SO_2 concentration. The fit is slightly worse in the summer season, when a certain number of zero pollution hours occour.

Spectral analysis

Auto-spectra of concentrations, wind speed, temperature and pressure as well as the amplitude, phase and coherence parameters of the corresponding cross-spectra have been evaluated. This spectral analysis does not increase significantly the general information on the phenomenon. Concentration series versus meteorological series show higher level of coherence in correspondence with diurnal and semi-diurnal oscillations, but there is no definite oscillation of period greater than one day. However, almost all the stations show a certain increase of coherence, between SO_2 and meteorological data, in correspondence with periods in the range 2.5-3.5 days.

Box-Jenkins ARIMA models

ARIMA models and seasonal ARIMA models have been applied to both 30-minute SO_2 measurements (Stations 3,4,5,8,10,11), and to 4-hour, daily, weekly average SO_2 values (Station 3) during the period of analysis (September 1974-October 1975). The application of several models of this type has led to the following general conclusion: AR(1) model (first order autoregressive process) shows good forecasting performances and, in any case, an efficiency comparable with that of more complex ARIMA models. Stationary AR(1) process can be expressed as

$$x_t - \mu = \varphi (x_{t-1} - \mu) + \varepsilon_t \qquad (1)$$

where x_t is the average SO_2 concentration in the t-th time step; $\mu = E\{x_t\}$ (E = expectation); $\{\varepsilon_t\}_t$ = white noise; and φ is the model parameter.

30-minute average The forecasting of next 30-minute SO_2 averages on the basis of previously observed pollution levels has given the satisfactory results summarized in Table 1 (columns 1 to 6), in which AR(1) model forecasting performances (column 6) are compared with those of the persistence model (next 30-minute average is equal to previous 30-minute average, column 5). Results show the general improvement obtained with AR(1) predictors.

Table 1 For each station and for two seasons are shown the average SO_2 value, the variance of the data, the AR(1) model parameter, the noise variance of: persistence model, AR(1) model, and cyclostationary AR(1) model.

Station number		μ	Variance of the data	AR(1) φ	σ_ε^2 variance of white noise		
					Persistence model	AR(1)	AR(1)CS
winter 74-75	03	97	6241	0.83	2104	1927	1693
	04	87	7609	0.78	3313	2954	2772
	05	52	3613	0.89	808	763	711
	08	86	6866	0.83	2346	2146	1932
	10	22	1913	0.89	417	394	354
	11	74	4391	0.81	1638	1486	1358
summer 75	03	97	22686	0.86	6234	5806	5251
	04	14	439	0.74	232	201	176
	05	--	-----	----	----	----	----
	08	14	514	0.72	291	250	226
	10	40	2962	0.70	1763	1501	1284
	11	27	1426	0.75	723	631	563

4-hour, daily and weekly average The most polluted station in the area (Station 3) has been used as a "pilot" station for an accurate analysis of average SO_2 pollution levels. Table 2 (columns 1 to 6) shows the efficient application of AR(1) predictor, with the exception of the weekly averages which present a non-correlated behaviour ($\varphi = 0.02$).

Cyclostationary predictor models
The presence of a daily cycle in the SO_2 measured data, has suggested to apply to cyclostationary (CS) processes (Franks,

Table 2 4-hour, daily and weekly average SO_2 values measured at Station 3 during the period November 1974-October 1975. The Table shows the variance of the data, the AR(1) model parameter, the noise variance of: persistence model, AR(1) model, and cyclostationary AR(1) model.

Averaging time	μ	Variance of the data	AR(1) φ	σ_ε^2 variance of the white noise		
				Persistence model	AR(1)	AR(1)CS
4-hour	94	13362	0.55	12124	9375	8237
daily	94	6115	0.53	5768	4413	----
weekly	94	2634	0.02	5256	2683	----

1971). Precisely, the AR(1)CS process has been considered, namely the model described by

$$\tilde{x}_{i+1}(k) = \varphi_i \, \tilde{x}_i(k) + \varepsilon_i(k) \qquad (2)$$

where $\tilde{x}_i(k) = (x_i(k) - \mu_i) / \sigma_i$; $x_i(k)$ is the average concentration in the i-th time step of the k-th day (the day is divided in N time steps, i=1,2,...,N); $\mu_i = \sum_{k=1}^{K} x_i(k) / K$ and $\sigma_i = \left\{ \sum_{k=1}^{K} (x_i(k) - \mu_i)^2 / K \right\}^{1/2}$ are the average and the standard deviation at the i-th time step of the day (period of analysis = K days); $\left\{ \varepsilon_i(k) \right\}_{i,k}$ = white noise; $\left\{ \varphi_i \right\}_i$ = set of model parameters; and $\tilde{x}_{N+1}(k) = \tilde{x}_1(k+1)$. With respect to the variance of the white noise, the cyclostationary model exhibits a substantial improvement if compared with the AR(1) (Table 1 and Table 2, last column). In particular, it must be noticed that the cyclostationary models take into account the daily cycle better than the best representative of the seasonal ARIMA models, which, for thirty-minute averages, can be described in the following way

$$\nabla^2 \nabla_{48} x_t = (1 - \vartheta_1 B) (1 - \Theta_1 B^{48}) \varepsilon_t \qquad (3)$$

where $\nabla x_t = x_t - x_{t-1}$; $\nabla_{48} x_t = x_t - x_{t-48}$; $B \varepsilon_t = \varepsilon_{t-1}$; $B^{48} \varepsilon_t = \varepsilon_{t-48}$; $\left\{ \varepsilon_t \right\}_t$ = white noise ; and ϑ_1 , Θ_1 are the model parameters.

Adaptive models

All the above discussed predictors show a general decrease of forecasting accuracy when used during periods different from those in which their parameters were calibrated. This is a typical behaviour of statistical models. In order to remove this shortcoming, an application has been carried out of adaptive AR(1) and AR(1)CS models, whose parameters are estimated on a learning past period of given fixed length immediately preceding the forecasting time. In this way the learning period moves along the SO_2 time series. The adaptive models can be used to represent the entire process, without distinguishing from season to season, according to their adaptive ability. Application of such a method allows to use AR(1) predictors without distinguishing between "data fitting" (i.e., forecasting in the same period used for parameters estimation) and "real time forecasting". Neverthless, they provide results comparable with those shown in Table 1 and Table 2.

Multiple regression models

In order to improve the SO_2 levels forecasting performances, it has been taken into account the meteorological input by using two simple regression models including SO_2 average data and the following meteorological parameters: temperature, wind speed and percentage of polluting wind directions occurrence. Daily and 4-hour average SO_2 values at Station 3, and the correspondent meteorological measurements, have been obtained by applying the following models:

model I : $$C_t = a\ T_t + b\ P_t + c\ S_t + d + \varepsilon_t \qquad (4)$$

model II : $$C_t = a'\ T_t + b'\ (P_t\ /\ S_t) + c' + \varepsilon'_t \qquad (5)$$

where C_t = average SO_2 concentration; T_t = average temperature; P_t = percentage of polluting winds (from S and SE for Station 3); S_t = average wind speed (all the above defined in the t-th time step); a, b, c, d, a', b', c' are the models parameters; and ε_t , ε'_t are the errors of the models in the t-th time step. By least squares estimation of the parameters it is possible to obtain the series $\left|\varepsilon_t\right|_t$, $\left|\varepsilon'_t\right|_t$ which can be described with an AR(1) model in order to decrease the variance of the white noise. By this separation between the "meteorological" and the "stochastic" contribution a predictor has been obtained which can be used as a first alert tool in forecasting pollution episodes. Computation of model parameters shows the validity of the approach according to the physics of transport and diffusion of pollutants (b and $b' > 0$, and $c < 0$). Table 3 shows the forecasting improvement obtained especially by applying model I (Equation 4). Figures 2 and 3 show the winter and the summer measured and forecast (model I) daily SO_2 average values. These models have been applied by using the measured meteorological parameters which, in general, are unknown in advance. In this way models I and II show, in Table 3 and in Figures 2 and 3, their best "theoretical" performances, although an efficient use of this methodology needs an estimation of the meteorological trend. However, the above results show the potential

Table 3 Analysis of 4-hour and daily average SO_2 values. Comparison of the variances of the white noise of the persistence model, AR(1) model, model I and model II defined by Equations 4 and 5.

	Season	μ	Variance of the data	σ_ε^2 variance of the white noise			
				Persistence model	AR(1)	model I	model II
4-hour	winter 74-75	97	4111	2799	2325	2168	2344
	summer 75	97	16284	17348	12736	7647	11129
daily	winter 74-75	97	2487	1929	1546	1017	1235
	summer 75	97	5144	7947	4939	2430	2762

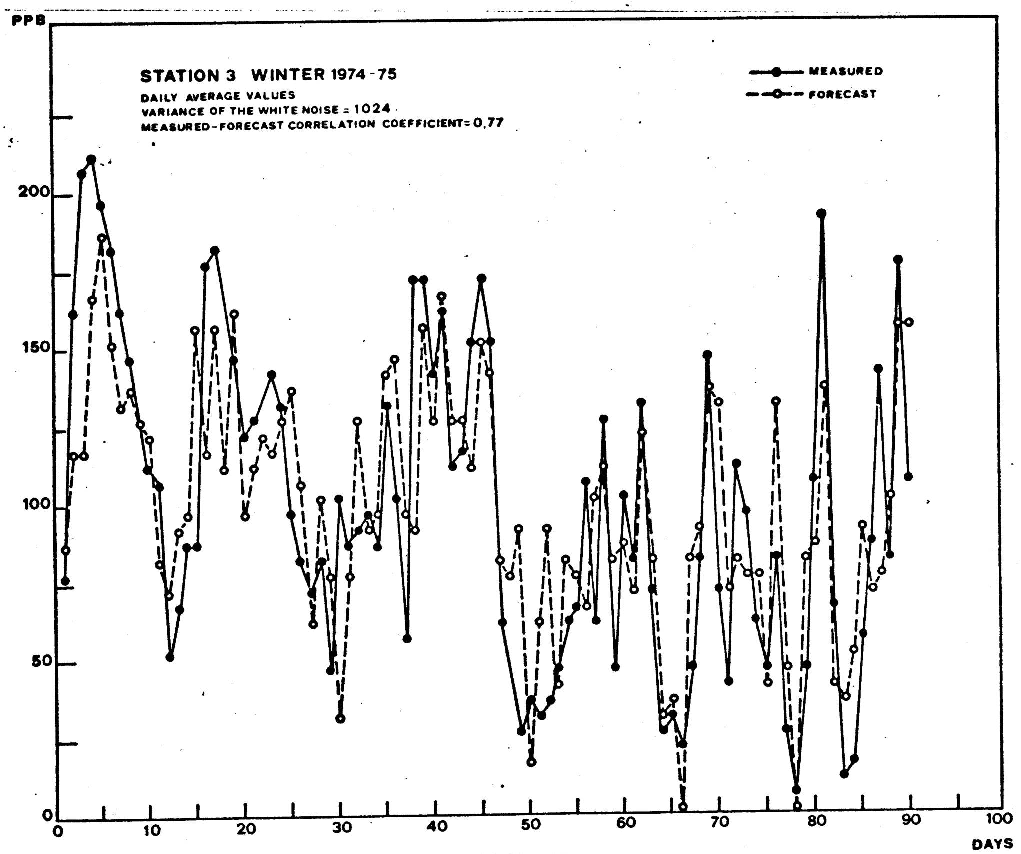

Figure 2 Measured and forecast (model I) daily SO_2 average values during the winter 1974-75 (correlation coefficient of the persistence model = 0.60).

efficiency and the adaptive ability of this approach, and suggest to develop this methodology.

DPP pollution index

In order to take into account the whole pollution phenomenon in the Venetian area, without using generally complex multivariate stochastic models and predictors, a meaningful air pollution index has been defined as a global representation of the average damage on the individual living in the area. This index, called "dosage population product" or DPP, takes the three following factors into account: 1) the pollution level; 2) the duration of pollution events; and 3) the number of people exposed to pollution (Finzi et al., in press). In fact, DPP in the k-th day is defined as

$$DPP(k) = \sum_{i=1}^{N} P_i \, D_i(k) \; / \; P \qquad (6)$$

where P_i is the number of people living in the i-th district of the city; P is the total population ($P = \sum_{i=1}^{N} P_i$) ; N is the

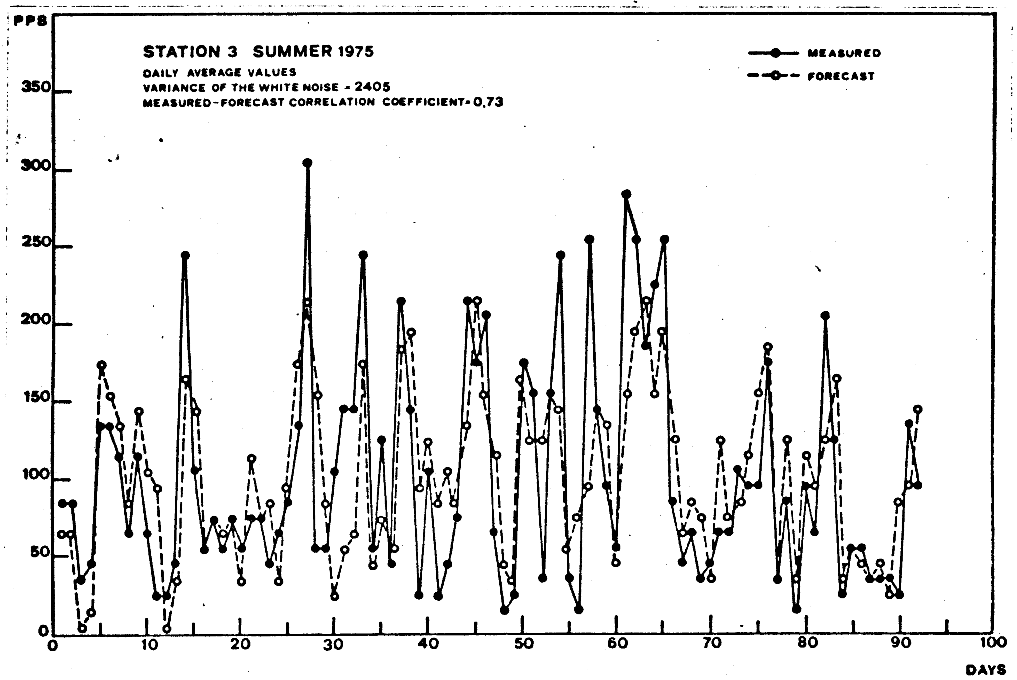

Figure 3 Measured and forecast (model I) daily SO_2 average values during the summer 1975 (correlation coefficient of the persistence model = 0.24).

number of population districts; and $D_i(k)$ is the k-th dosage (integral of SO_2 concentration over the k-th day) in the center of the i-th district. Several stochastic models have been applied in order to forecast the daily "observed" DPP. Best predictors show a correlation coefficient of 0.6 which can constitute a first important result in order to control the effective damage suffered by population living in the study area.

CONCLUSIONS

The statistical approach to air pollution study has shown its validity in time series analysis and forecasting. Applications of statistical models to Venice meteorological and SO_2 data allows a quantitative control of the different performances of these models. However, it must be pointed out both the advantages and the limitations of this methodology (Myrabo et al., 1975). Statistical/empirical models have two major advantages. The first is their close relationship to the actual atmospheric data. Thus one can hope to predict successfully even when deterministic understanding of the complex real world is incomplete. The second advantage is the relative simplicity and low cost of the development and use of statistical/empirical models. On the other side, this approach shows the disadvantage that one is not assured of reliability in extrapolating the model beyond the range of conditions contained in the data from which it was derived. In addition to that, it must be pointed out that, in short term air quality forecasting, one should not expect to be able to predict air quality better than the weather.

Acknowledgements

The author is indebted with Ente Zona Industriale di Porto Marghera for the meteorological and SO_2 data supplied. A sincere appreciation is expressed to Dr. S. Boscolo and Mr. M. Fantini of the Venice IBM Scientific Center for their help in the data base arrangement.

References

Bencala, K.E. and Seinfeld, J.K. (1976) On frequency distributions of air pollutant concentrations. Atmospheric Environment, Pergamon Press, 10, 941-950.

Box, G.E.P. and Jenkins, G.M. (1970) Time series analysis forecasting and control. Holden-Day, San Francisco.

Chock, D.P., Terrel, T.R. and Levitt, S.B. (1974) Time-series analysis of Riverside, California air quality data. Atmospheric Environment, Pergamon Press, 9, 978-989.

Finzi, G., Fronza, G., Rinaldi, S. and Zannetti, P. (in press) Modelling and forecast of the dosage population product in Venice. IFAC Symposium on Environmental System Planning Design and Control, Kyoto, Japan.

Franks, L. (1971) Signal Theory. Prentice Hall, Henglewood Cliffs, New Jersey.

Hunt Jr., W.F. (1972) The precision associated with the sampling frequency of log-normally distributed air pollution measurements. Journal APCA, 22, 9:687-691.

Kahn, H.D. (1973) Note on the distribution of air pollutants. Journal APCA, 23, 11:973.

Larsen, R.I. (1969) A new mathematical model of air pollutant concentration averaging time and frequency. Journal APCA, 19, 1:24-30.

McCollister, G.M. and Wilson, K.R. (1975) Linear stochastic models for forecasting daily maxima and hourly concentrations of air pollutants. Atmospheric Environment, Pergamon Press, 9, 417-423.

Mertz, P.H., Painter, L.J. and Ryasón, P.R. (1972) Aerometric data analysis-time series analysis and forecast and an atmospheric smog diagram. Atmospheric Environment, Pergamon Press, 6, 319-342.

Myrabo, L.N., Wilson, K.R. and Trijonis, J.C. (1975) Survey of statistical models for oxidant air quality prediction. Conference on State-Of-The-Art of Assessing Transportation Related Air Quality Impacts, Washington, D.C..

Runca, E. and Zannetti, P. (1973) A preliminary investigation of the air pollution problem in the Venetian area. Technical Report, IBM Scientific Center, Venice.

Runca, E., Melli, P. and Zannetti, P. (1976-A) An application of air pollution models to the Venetian area. Seminar on Air

Pollution Modelling, IBM Scientific Center, Venice.
Runca, E., Melli, P. and Zannetti, P. (1976-B) Computation long-term average SO_2 concentration in the Venetian area. Applied Mathematical Modelling, 1, 9-15.

Tiao, G.C., Phadke, M.S. and Box, G.E.P. (1976) Some empirical models for the Los Angeles photochemical smog data. Journal APCA, 26, 5:485-490.
Tilley, M.A. and McBean, G.A. (1973) An application of spectrum analysis to synoptic-pollution data. Atmospheric Environment, Pergamon Press, 7, 793-801.
Trivikrama Rao, S., Samson, P.J. and Pedadda, A.R. (1976) Spectral analysis approach to the dynamics of air pollutants. Atmospheric Environment, Pergamon Press, 10, 375-379.
Zannetti, P., Melli, P. and Runca, E. (in press) Meteorological factors affecting SO_2-pollution levels in Venice. Atmospheric Environment, Pergamon Press.
Zannetti, P. (in press) Time series analysis of Venice air quality data. IFAC Symposium on Environmental System Planning Design and Control, Kyoto, Japan.