Time-series analysis of air pollution data

Article *in* Atmospheric Environment · July 1999

DOI: 10.1016/S1352-2310(99)80001-6

4 authors:

Romualdo L. Salcedo
University of Porto

Monaliza Antunes Ferraz
Federal University of Minas Gerais

Célia Alves
University of Aveiro

Fernando Gomes Martins
University of Porto

PERGAMON

Atmospheric Environment 33 (1999) 2361–2372

ATMOSPHERIC ENVIRONMENT

Time-series analysis of air pollution data

R.L.R. Salcedo[a,]*, M.C.M. Alvim Ferraz[a], C.A. Alves[b], F.G. Martins[a]

[a] *Departamento de Engenharia Química, Faculdade de Engenharia da Universidade do Porto, R. dos Bragas, 4099 PORTO CODEX, Portugal*

[b] *Departamento de Ambiente da Universidade de Aveiro, 3810 AVEIRO, Portugal*

Received 17 March 1998; received in revised form 28 October 1998; accepted 11 November 1998

Abstract

Time-series analysis of air pollution environmental levels involves the identification of long-term variation in the mean (trend) and of cyclical or periodic components. A model based on a stepwise approach to time-series analysis was applied to the daily average concentrations of strong acidity (SA) and black smoke (BS) in the Oporto area, using an available computer program. Each step is completed by a correlation analysis of the residuals, allowing the identification of an optimal structure with a residual white noise. A periodic component with harmonics defined through "peaks" of concentration on week middle days and "troughs" on weekends was observed. SA concentration behaviour can be related with industrial activities, mainly through fossil-fuel burning in discontinuous working cycles. The observed evolution for BS is most probably related with weekly patterns of motor traffic, with observed minimum values during weekends. The periodic components represent, on the average, about 5% of the total variance for the SA series and 15% for the BS series. However, the weekly cycles are predominant in the SA series, representing on the average 75% of the periodic variance, against 46% for the BS series. Statistically significant higher frequency ($\approx$2–4 day) periodic components were observed for both pollutant indicators and for all collection sites analysed. This may be due to synoptic weather variations of minimum and maximum daily temperature and precipitation, which show similar periods in the study area.

Keywords: Time series; Strong acidity; Black smoke; Air quality

Nomenclature

x_t	24 h-average observation at day t ($\mu g\,m^{-3}$)
y_t	variance stabilised series (ex. $y_t = \log(x_t)$)
c_k	autocorrelation coefficient in time domain ($\mu g\,m^{-3})^2$
N	number of observations
k	lag (days)
μ	average for entire series ($\mu g\,m^{-3}$)
V_k	cosine autocovariance in the frequency domain ($\mu g\,m^{-3})^2$
L	maximum lag (days)
m_t	trend at day t ($\mu g\,m^{-3}$)
s_t	periodic component at day t ($\mu g\,m^{-3}$)
e_t	error component (residual) at day t ($\mu g\,m^{-3}$)
a_p	least-squares cosine coefficient of Fourier series ($\mu g\,m^{-3}$)
b_p	least-squares sine coefficient of Fourier series ($\mu g\,m^{-3}$)
p	harmonic number
r	correlation coefficient
R_p	variance of pth harmonic ($\mu g\,m^{-3})^2$

* Corresponding author. E-mail: rsalcedo@fe.up.pt.

1352-2310/99/$ - see front matter
PII: S1352-2310(98)00407-5

Acronyms

BS – black smoke
SA – strong acidity
SATSA – sequential approach to time series analysis

1. Introduction

A time series is a record of observations made at a particular location, associated with the evolution with time of some particular variable. Time series occur in a great variety of phenomena, ranging from economics to the physical sciences, and general methodologies may be applied without any need for prior knowledge of the underlying causal links between the dependent and independent variables (viz., a black-box approach). However, this missing causality link introduces an inherent uncertainty in the fitting and especially in the forecasting estimates. Also, the observed behaviour cannot be replicated with repeated experiments (a single realisation occurs at each point in time) and the observations are usually dependent over time.

Air quality measurements constitute good examples of environmental time series. The methodology which is usually employed for the estimation of environmental parameters is based on classical descriptive statistics, but this is of a rather limited value due to the large variability associated with air quality data and the low signal-to-noise ratio of the available measurements. Time-series analysis may be a good approach to avoid these difficulties by allowing the identification of hidden deterministic behaviour and thus contributing to an understanding of cause and effect relationships in environmental problems (Schwartz and Marcus, 1990).

Time-series analysis may be broadly divided into purely statistical methods applicable to non-repeatable experiments (Box and Jenkins, 1970; Robinson and Silvia, 1979; Hoff, 1983; Montgomery et al., 1990) and more or less complex structural models (Chatfield, 1989; Montgomery et al., 1990; Kendall and Ord, 1990; Schlink et al., 1997). The Box–Jenkins approach has been thoroughly applied to the analysis of a variety of time series from the social and economics' sciences. The application to physical sciences and environmental series is also described in the literature (Chock et al., 1975; Simpson and Layton, 1983; Crabtree et al., 1990; Zannetti et al., 1990; Milionis and Davies, 1994a, b). Basically, the Box–Jenkins approach uses the concepts of autoregression (AR) and moving average (MA), where the dependent variable under study is lag-regressed onto itself and smoothed, thus giving rise to the so-called ARMA and related ARIMA and SARIMA models (integrated and seasonal integrated models). These models are applicable to (at least weakly or second-order) stationary series, where there is no systematic change in mean (i.e., the series has been detrended) and the variance is constant over time (Chatfield, 1989; Kendall and Ord, 1990). Structural models, on the other hand, attempt to decompose the time series into individual components, which are usually a trend (long-term variation in the mean), cyclical components (seasonality and other cyclical or quasi-cyclical terms) and an unexplained random residual (independent and normally distributed, viz., a white noise). Although the application of structural models is uncommon in atmospheric sciences, spectral analysis with recursive Kalman filters have been recently applied to SO_2 air pollution levels (Schlink et al., 1997).

These two different approaches are perhaps better described as complementary, because while ARIMA modelling offers the advantage of a stronger model building paradigm, structural models offer clearer interpretations through the decomposition into components (Kendall and Ord, 1990). The choice of an appropriate Box–Jenkins model is also not straightforward, requiring a rather elaborate procedure to identify and validate an adequate model, and may be somewhat subjective (Chatfield, 1989), depending mainly on the objectives of the analysis, with the possibility that a wrong model may be chosen, since different models may actually describe very similar time series (Milionis and Davies, 1994a). The application of structural models is also non-trivial, since there is some subjective interpretation of correlograms and spectra which cannot be avoided but only minimised. The structural approach also requires that the series be stationary in the mean, and incorrect detrending methods may seriously distort the variance spectra and autocorrelation functions (Kendall and Ord, 1990).

Relative to their applicability to ambient air pollution time series, it is rather unlikely that either ARIMA or structural models may show any reasonable predictive capabilities beyond the detection of trends and cyclical components (Benarie, 1980, 1987). However, Schlink et al. (1997) have built a structural model based on a recursive Kalman filter that gives estimates of hourly SO_2 concentrations in good agreement with measured values. Also, even though well-defined cyclical components may be present, these may not be strictly periodic (Schlink et al., 1997), and extrapolation under such circumstances should be avoided. The detection of cyclical components in urban air pollution, other than the obvious weekly cycle, may however be used as a tool for source apportionment, since the frequency and amplitude characterisation of these cyclical components may aid in the identification of the offending source(s).

A computer program (SATSA, for sequential approach to time series analysis) was developed based on a stepwise structural approach to time series, whereby trend, cyclical components and noise are obtained through an appropriate observation equation (Salcedo and Dias, 1992). The model does not assume any a priori knowledge of the noise level and is applied in a sequential mode with the objective of obtaining at each step a residual which approaches a white noise. Normality and independence tests are applied to the residuals as each harmonic is identified (through Fourier series or fast Fourier transforms) to minimise the occurrence of periodic noise which may otherwise be artificially induced by the harmonic decomposition. The program incorporates logarithmic and square root data transformations for variance stabilisation and first- and second-order differencing for trend removal. Autocorrelation and smoothed variance spectra are also available as a Fourier preprocessing step and residuals' checking. The usual dependence of Fourier harmonic decomposition on the time series length is avoided through a simple optimisation procedure, viz. a frequency fine tuning, and filtering is applied to ensure physically meaningful results. The SATSA program was tested with a variety of time series described in the literature, taken both from the econometrics and physical sciences, showing robust fitting and forecasting capabilities (Salcedo and Dias, 1992).

In this work, to exemplify the model capabilities, the SATSA program was applied to a variety of time series of strong acidity (SA) and black smoke (BS) available for the Oporto Metropolitan Area, showing the occurrence of well-defined short-term cyclical behaviour despite the prevalence of random components. The composite periodic components for all analysed series have "peaks" on weekdays and "troughs" on weekends, irrespective of their industrial and/or urban characteristics. Apart from a well-defined weekly cycle, high-frequency ($\approx$2.2–4.0 days) components were identified in all pollution series. Since similar high-frequency components ($\approx$2.7–4.5 days) are observed for records of minimum and maximum daily temperatures and precipitation for the study area, this may be an indication of a synoptic weather causal effect on air pollution levels.

2. The SATSA model

The SATSA program is written in Fortran 77 for an easy portability between different computers, and a brief description follows. A time series of air quality data may be represented through one of the following observation equations (Rich, 1973; Chatfield, 1989; Kendall and Ord, 1990):

$$x_t = m_t + s_t + e_t, \tag{1}$$

$$x_t = m_t\, s_t\, e_t, \tag{2}$$

where x_t is the observed value at time t, and m_t, s_t and e_t are, respectively, the trend, cyclical and error components, viz. the components to be identified. Eq. (1) corresponds to an additive model, whereas Eq. (2) is a multiplicative one, which can be converted into an additive model through a logarithmic transformation. Other models, which are not considered in the SATSA program, may, however, be useful for particular situations, such as the multiplicative-seasonal model (Kendall and Ord, 1990). In order to identify the different components, the SATSA program performs a series of tasks described below.

2.1. Detrending analysis

Since a stationary series in the mean is a prerequisite to a correct frequency spectrum estimate, detrending the series is a very important operation that should be correctly performed. There are several methods usually employed, ranging from moving averages to polynomial regressions to differencing. Since it is known that moving averages may introduce autocorrelation between the residuals, the so-called Slutzky–Yule effect (Yule, 1927; Slutzky, 1937; Kendall and Ord, 1990), detrending should not be performed with moving averages, although some authors do so (Dejak et al., 1990). The SATSA program allows detrending via linear regression or first- or second-order differencing, defined, respectively, as

$$\nabla x_{t+1} = x_{t+1} - x_t, \tag{3}$$

$$\nabla^2 x_{t+2} = \nabla x_{t+2} - \nabla x_{t+1} = x_{t+2} - 2x_{t+1} + x_t, \tag{4}$$

where the original series is substituted by the detrended series and all subsequent analyses are made in the modified series. There has been a recent tendency in the econometrics literature in favouring trend removal by differencing as opposed to polynomial regression (Kendall and Ord, 1990), and this practice has been retained below for the analysis of all the time series reported in the present work. The need for differencing may be estimated by visual inspection of the correlograms, but formal tests are available (Fuller, 1976). Although these tests are not currently included in the model, they were performed for all series studied in the present work. Basically, a linear regression is made between the variance-stabilised series y_t (see *Data transformation and filtering* below) on 1 and on y_{t-1}, for testing first-order differencing, and between ∇y_t on 1 and on ∇y_{t-1} for testing second-order differencing (Schlink et al., 1997). This is equivalent to performing a multiple linear regression of y_t on 1, y_{t-1} and ∇y_{t-1}. The regression coefficients are analysed for statistical significance, using appropriate empirical tables (Fuller, 1976).

Proper detrending cannot be overemphasised, since the nonstationarity inherent in many time series accounts for most of the spurious relationships apparently observed (Milionis and Davies, 1994a).

2.2. Autocorrelation and variance spectrum estimation

This module includes subroutines that act as a preprocessing step to the Fourier decomposition and as a verification step for the correlation between residuals. Basically, the autocorrelation function

$$c_k = \sum_{i=1}^{N-k} = [(x_i - \mu)(x_{i+k} - \mu)]/(N-k) \qquad (5)$$

gives, in the time domain, the degree of dependence between k-lagged values for the entire N-valued series. The SATSA program normalises the autocovariance coefficients c_k to the autocovariance coefficient at lag zero c_0, as is usually done (Rich, 1973; Payne, 1988; Chatfield, 1989). A complementary method of detecting correlation is through the Fourier transformation of the autocovariance function from the time domain to the frequency domain (Rich, 1973)

$$V_k = (C/L)[c_0 + c_L \cos(k\pi) + 2\sum_{r=1}^{L-1} c_r \cos(rk\pi/L)],$$
$$k = 0, L, \qquad (6)$$

where $C = 1$ for $k = \{1, \ldots, L-1\}$, $C = 0.5$ for $k = \{0, L\}$ and L is the maximum lag. It can be shown that the variance spectrum defined by Eq. (6) does not give consistent estimates of the spectrum (Rich, 1973; Chatfield, 1989; Kendall and Ord, 1990). To correct this problem, the SATSA program employs a smoothing procedure of the variance estimates in the frequency domain (the Hanning window).

2.3. Fourier analysis

The SATSA program performs a typical spectrum analysis through Fourier series decomposition into distinct harmonics, but adds an optimisation capability to the frequency estimation. The reason for this is to correct the dependence of the Fourier harmonic decomposition on the time-series length, since the harmonics have to be integer divisors of this length. If the correct frequency contributes to a large proportion of the total variance then the variance explained by an approximate frequency may be much smaller and the residuals will be in error. Fine tuning of the frequency is made by finding the neighbouring frequency that maximises the variance (Salcedo and Dias, 1992). The Fourier representation for the periodic component after detrending may be represented by (Payne, 1988; Chatfield, 1989; Kendall and Ord, 1990)

$$s_t = \sum_{p-1}^{m} [a_p \cos(2\pi pt/N) + b_p \sin(2\pi pt/N)], \qquad (7)$$

where the value of m depends on whether the series has an even or odd number of points. The coefficients of the Fourier series are obtained by least-squares as

$$a_p = 2\left[\sum_{i=1}^{N} x_i \cos(2\pi pt/N)\right]\Big/N; \quad p = 1, m, \qquad (8)$$

$$b_p = 2\left[\sum_{i=1}^{N} x_i \sin(2\pi pt/N)\right]\Big/N; \quad p = 1, m, \qquad (9)$$

and can be calculated faster through the use of fast-Fourier transforms (Brigham, 1974; Press et al., 1986). The contribution of the pth harmonic to the total variance (Parseval's theorem) is given by

$$R_p = (a_p^2 + b_p^2)/2. \qquad (10)$$

The Fourier series has N parameters to describe N observations, and so can be made to fit the data exactly (Chatfield, 1989; Kendall and Ord, 1990). Obviously, this does not mean that the observed series is cyclical, or even that any periodicity is indeed occurring. Thus, one has the problem of deciding how many and which harmonics are indeed significant (Rich, 1973; Minors and Waterhouse, 1988). This is a serious question which has to be carefully addressed, otherwise wrong conclusions and contamination of the residuals with periodic noise will inevitably occur (Kendall and Ord, 1990; Minors and Waterhouse, 1988).

2.4. Statistical significance and residual analysis

A statistical F ratio is computed using Whittle's extended Fisher test for each harmonic to assess its significance (Minors and Waterhouse, 1988; Murteira et al., 1993). The associated probability P-value is also computed (Gordon and Gordon, 1994).

It is assumed that true random residuals should be normally distributed and independent (Rich, 1973; Payne, 1988; Chatfield, 1989; Montgomery et al., 1990; Kendall and Ord, 1990; Milionis and Davies, 1994a). The residuals are thus analysed after each harmonic decomposition through normality and independence tests. In the normality test, the residuals are first sorted and classified as a probability distribution, and subsequently are linearly regressed on the inverse normal distribution that has the same mean and standard deviation as the observed distribution. The independence test is a simple turning points test (Kendall and Ord, 1990), where a turning point may either be a "peak" if it is greater than its two neighbours or a "trough" if it is lower than its two neighbours. For series with at least 100 observations the distribution of turning points is approximately normal, and a simple t-test can then be used to establish independence between the residuals. In the SATSA program residuals that show a number of turning points within 1% of the expected number for an independent distribution are considered to be independent.

Thus, the SATSA program starts the harmonic decomposition (after detrending) by removing the most significant harmonic. The residuals are then analysed for normality and independence and the process repeated. The decomposition ends when the residuals are independent and normally distributed or when the contribution of the current harmonic to the total variance drops below a pre-established very low value, viz. 0.1%.

2.5. Data transformation and filtering

This module performs logarithmic and square-root back-and-forth transformations for variance stabilisation and to transform a multiplicative model into an additive one (viz., Eq. (2) into Eq. (1)). To decide which transformation to make, the original series is divided into smaller subsets and the local mean, standard deviation and variance are computed. Linear regressions between the local means and standard deviations and variances are then performed. If the local mean is proportional to the local variance, then the variance is stabilised with a square-root transformation, and if it is found to be proportional to the local standard deviation, then the variance is stabilised with a logarithmic transformation (Chatfield, 1989; Milionis and Davies, 1994b). A simple filtering procedure (Bartnicki, 1989) is applied to the model data in order to remove negative values, for example when working with environmental time series (Schlink et al., 1997).

This module also converts the first- or second-order differenced data into the corresponding undetrended data. The SATSA model thus employs a decomposition approach based on the assessment of the statistical significance (F-ratios and P-values) of the distributed variance, on fine tuning of the identified periods and on residuals checking for normality and independence. These procedures distinguish this model from off-the-shelf currently available Fourier decomposition models (e.g., MATLAB).

3. Methodology of analysis

The present paper is based on results collected in the Air Quality Network of Oporto Area, under the respective Air Management Commission. Oporto is the second largest city in Portugal, located in the North (approximate latitude and longitude: 41° 10′ North and 8° 40′ East). The data was collected between March 1986 and March 1992. Strong acidity (SA) and black smoke (BS) were measured with sequential semi-automatic SF_8 equipment. Daily averages of SA expressed as SO_2 were measured using the hydrogen peroxide method, with the end-point detection by potentiometry. BS was measured with the shade-reflectance method. Although the data is rather old, it is actually the most recent data available for SA and BS for the Oporto region, since the eight measuring stations were dismantled in 1993. There are presently two measuring stations fitted with on-line automatic samplers connected to a central computer, but the sampling of one station is defective since it is located directly behind a wall that avoids the predominant winds from reaching the sampling point. The other station is correctly located but since it is the only station, it does not allow comparison with data from other sampling sites. Thus, we have chosen to compare data available for SA and BS for the older network. Table 1 shows the characteristics of each site selected for the present study and Fig. 1 shows the location of the sampling sites. SA and BS time series were analysed with the SATSA program for these sites, which are considered the most representative ones of the entire air pollution monitoring network (Alvim Ferraz and Faria Ferraz, 1988; Alvim Ferraz et al., 1988). The lengths of the SA series are comparable with series from other studies (Milionis and Davies, 1994b), although the lengths for the BS series are shorter.

The PR site was once considered as a reference site, since it was placed upstream from major pollution sources, both of industrial and urban (traffic) origin, despite its location near the Oporto international airport. However, pollution levels have been steadily climbing at this site since 1986, and presently it already suffers pollution from traffic and industrial activities. Unfortunately, air pollution data for this site is not considered reliable before 1986 (Alvim Ferraz and Faria Ferraz, 1988), thus comparison with a "background" site is not possible. The MT site is mainly influenced by industrial emissions, since for the spring–summer period it is downwind from

Table 1
Representative sites of the air quality network of Oporto Area

Site	Characteristics	Measured variables	Period
PR	Was reference now is industrial and urban	SA daily averages	22 Sep. 1987–19 Nov. 1989
MT	Industrial	SA daily averages	14 Jun. 1988–30 Nov. 1990
ML	Urban	BS daily averages	8 Oct. 1990–31 May 1991 23 Jul. 1991–31 Dec. 1991
BV	Industrial and urban	SA daily averages	1978–1980 1986–1988

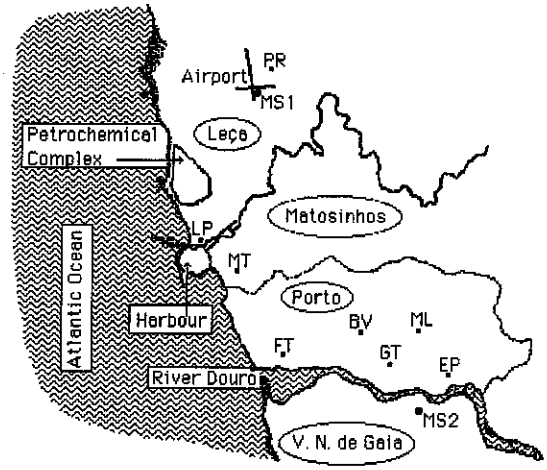


Fig. 1. Location of pollutant sampling sites and meteorological stations.

a major petrochemical complex and near the predicted location of the maximum concentration isopleths (Salcedo, 1990). The ML is the most representative site for traffic pollution. Thus, BS data were analysed with the SATSA program for this last site. The BV site is both influenced by traffic and industrial emissions, and was studied previously (Salcedo and Dias, 1992). It is included here only for comparison with the other sites.

Some of the series have gaps of less than seven days; in this situation, the SATSA program replaces the missing values by the linear trend of the available measurements. However, for the ML site the data was divided into two distinct time series, due to the occurrence of a significant number of consecutive days (54) without validated data.

The need to stabilise the variance was found by decomposing the various time series into smaller subsets, as explained above, to find whether significant correlations between the local means and variance or standard deviations occurred. In all cases, a logarithmic transformation was indicated. However, exactly the same results (periodic components and trend) are obtained if no transformation is performed, in agreement with the remarks of Nelson and Granger (1979) and Chatfield (1979). Detrending was performed through first-order differencing, which was found to be appropriate both from visual observation of the correlograms and from the application of Fuller's test at the 1% level of significance (Fuller, 1976).

4. Results and discussion

4.1. Time-series decomposition

Fig. 2a shows the original time series of SA daily means at the PR site, which has 790 data points. The first data point corresponds to 22 September 1987 (Tuesday), and the last one to 30 October 1990 (Monday). The data are clearly fluctuating and apparently random, with a long-term component corresponding to 19.5% of the total variance, and with some very high values. However, it is obvious that the long-term component is highly influenced by the length of the series, and thus should be regarded with great caution.

Fig. 2b shows the autocorrelation functions for the detrended data, as well as for the corresponding residuals obtained from the sequential decomposition. The autocorrelation function for the detrended values oscillates around zero, showing, however, some significant periodicities (Chatfield, 1989). The residuals autocorrelation function is similar to the original detrended autocorrelation function, a clear sign that the series is mostly random.

Fig. 2c shows the smoothed variance spectra for the original detrended and residual series. The original data shows statistically significant peaks (P-value below 0.05) at periods of 6.99, 2.34 and 2.80 days, which, respectively, explain 3.31, 1.19 and 0.45% of the total variance. This figure also shows that the variance spectrum is flatter for the residuals, due to the removal of the statistically significant cyclical components. The remaining peaks do not have any statistical meaning, e.g., they must be considered random. The residuals after decomposition are independent and no further information seems extractable from this set. The application of spectral analysis through FFT using the MATLAB Signal Processing Toolbox (Little and Shure, 1993) produced essentially the same periodic components (Fig. 2d), as long as the series had been previously differenced, and a visual identification of the three significant peaks is easily performed in this case.

Fig. 2e and f show the trend and periodic components superimposed on the raw data, for two different time periods. In Fig. 2e the first day corresponds to 22 September 1987 (Tuesday), and in Fig. 2f to 1 January 1988 (Tuesday). Thus, both series have "peaks" on weekdays (Wednesdays – Thursdays) and "troughs" on weekends. This cyclical behaviour is typical of the entire PR series, since it occurs if sub-series are analysed. Seasonal variations, such as detected by others, albeit for other pollutants (Derwent et al., 1998) were not detected in this series, perhaps because of a too short length.

Similar decompositions have been applied to the SA series obtained at the MT site, which has 900 data points, and very similar results have been obtained. The series shows statistically significant peaks at periods of 7.00, 4.03 and 2.30 days, which, respectively, explain 4.02, 0.68 and 0.23% of the total variance. Again, "peaks" of concentration appear on weekdays (Tuesdays–Wednesdays) and "troughs" on weekends. This weekly pattern must be due to anthropogenic activity, and has been found for other pollutants by Simmonds and Keay (1997).

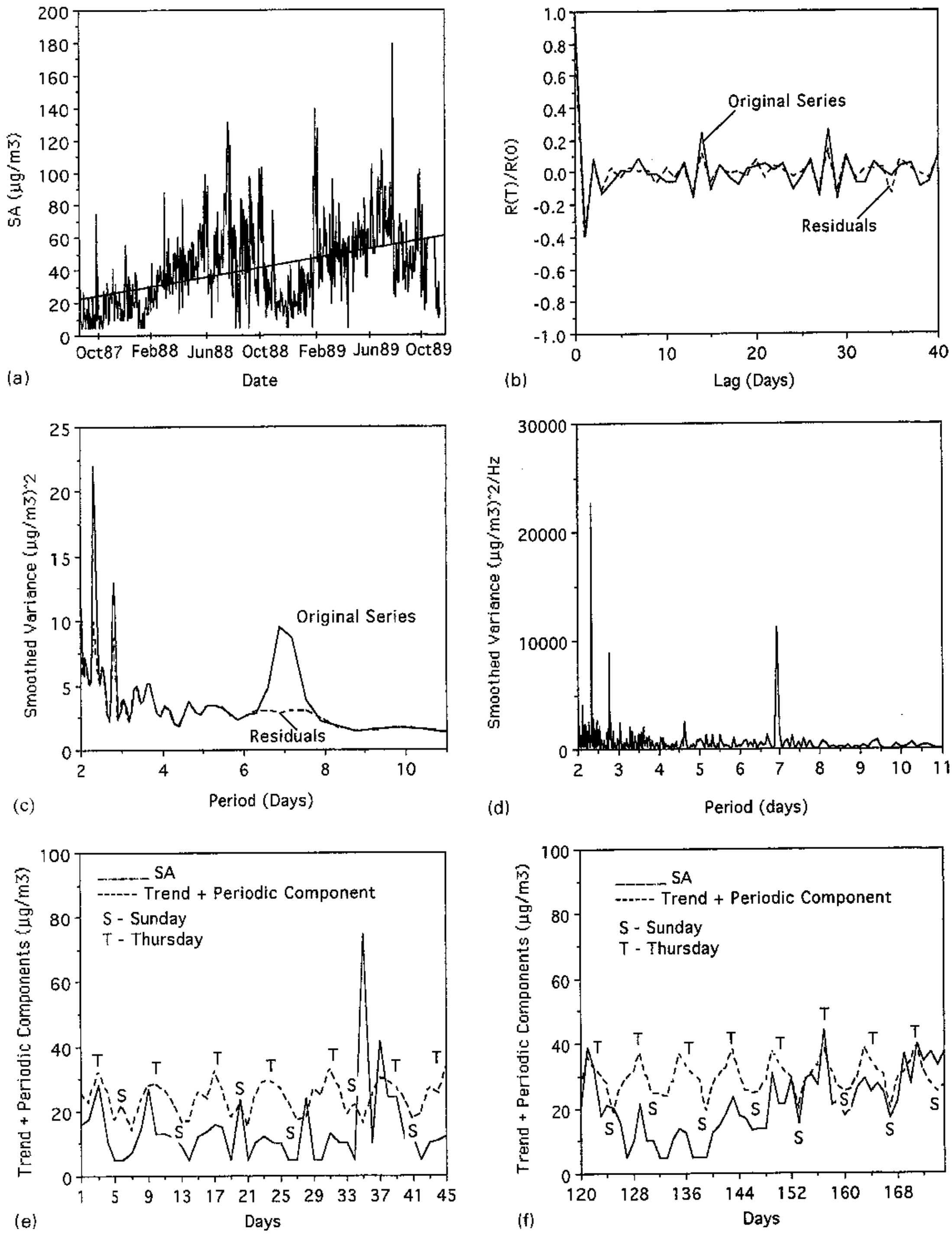


Fig. 2. a. Daily averages for SA at the PR site (1987–1989). b. Autocorrelation functions for the SA daily averages at the PR site (1987–1989). c. Smoothed variance spectra for the detrended and residual series – SA daily averages at the PR site (1987–1989). d. Smoothed variance spectra for the detrended series using the MATLAB Signal Processing Toolbox – SA daily averages at the PR site (1987–1989). e. Trend and periodic components – SA daily averages at the PR site (1987). f. Trend and periodic components – SA daily averages at the PR site (1988).

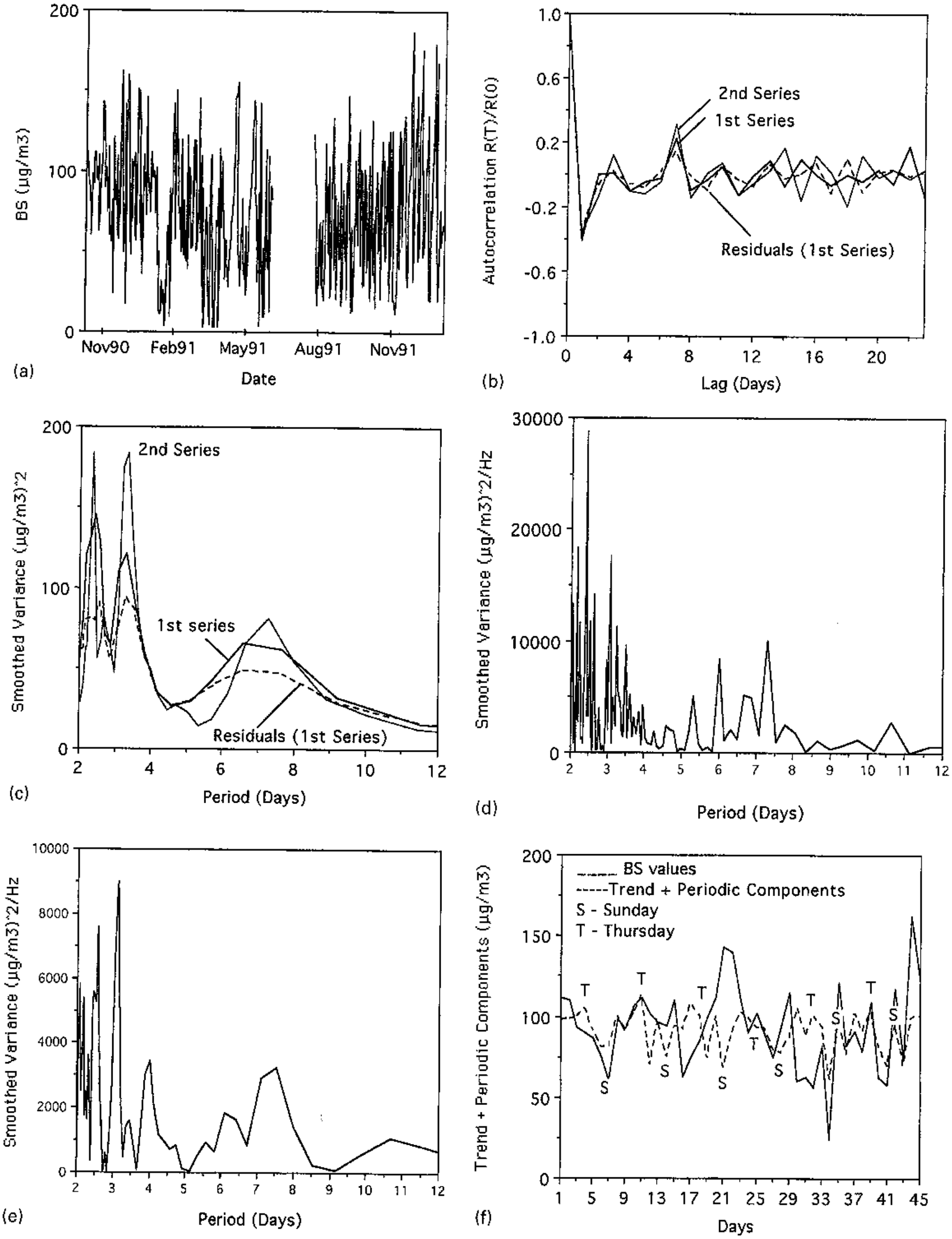


Fig. 3. a. Daily averages for BS at the ML site (1990–1991). b. Autocorrelation functions for the BS daily averages at the ML site (1990–1991). c. Smoothed variance spectra for the detrended and residual series – BS daily averages at the ML site (1990–1991). d. Smoothed variance spectra for the detrended series using the MATLAB Signal Processing Toolbox – BS daily averages at the ML site (1990–1991). e. Smoothed variance spectra for the detrended series using the MATLAB Signal Processing Toolbox with Welch's smoothing algorithm – BS daily averages at the ML site (1990–1991). f. Trend and periodic components – BS daily averages at the ML site (1990). g. Trend and periodic components – BS daily averages at the ML site (1990).

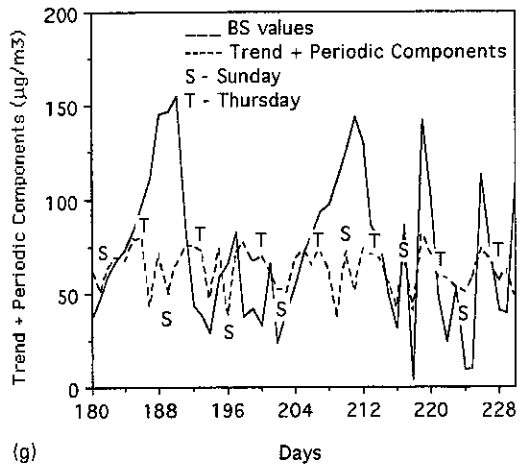


Fig. 3. (*continued*).

Seasonal variations could also not be detected in this series. Fig. 3a shows the original time series of BS daily means at the ML site, corresponding to both sub-series, which have, respectively, 235 and 162 data points. For the first sub-series, the first data point corresponds to 8 October 1990 (Monday), and the last one to 31 May 1991 (Friday). For the last sub-series, these are, respectively, 23 July 1991 and 31 December 1991, both a Tuesday. The data show very large fluctuations and a decreasing long term component for the first series (7.64% of the series variance) and an increasing one for the last series (5.94% of the series variance). The overall series shows almost no long term component.

Fig. 3b shows the autocorrelation functions for the detrended data, as well as for the corresponding residuals obtained from the sequential decomposition. The data corresponding to both series is very similar.

Fig. 3c shows the corresponding smoothed variance spectra, with statistically significant peaks for the first series at periods of 6.99, 2.43, 3.12 and 2.19 days, which, respectively, explain 4.06, 2.46, 2.04 and 1.63% of the total variance. For the second series, the variance spectrum is very similar, showing significant peaks at 7.02, 2.35 and 3.25 days, which account, respectively, for 11.10, 6.37 and 3.85% of the series variance. These proportions are much larger than those observed with the SA series, which may be due to the fact that BS values are more representative of local pollution, due to traffic, than SA values, which may have important contributions from more distant sources. The application of the MATLAB Signal Processing Toolbox to these series (Little and Shure, 1993) shows a highly confused spectrum (Fig. 3d), unlike the much cleaner spectrum from the SATSA model, and it is clearly very difficult to visually identify the significant peaks. Since a better estimate of the power spectrum may be obtained with MATLAB using Welch's method (Oppenheimer and Shafer, 1975), which involves averaging across adjacent records, this was done and the corresponding spectrum for the longer series can be seen in Fig. 3e. This spectrum clearly approaches that given by the SATSA model (Fig. 3c), although it shows a peak around 4 days that is not significant in this model.

Fig. 3f and g show the trend and periodic components superimposed on the raw data, for two different time periods. In Fig. 3f the first day corresponds to 8 October 1990 (Monday), and in Fig. 3g to 5 April 1991 (Friday). Just as it happened with the SA series, these BS series have "peaks" on weekdays (Wednesdays–Thursdays) and "troughs" on weekends.

It is interesting to compare these results with those previously obtained with two completely different data sets, also in the Oporto Area, but now for a different site (the BV site in Table 1) spanning each one two complete years (Salcedo and Dias, 1992). At the BV site, significant periods of 7.0 and 3.5 days were detected. Similar 3.5 day periods have been observed for time-series analyses performed in data collected in the northeastern United States (Tilley and McBean, 1973; Trivikrama et al., 1976), which is attributed to synoptic weather variations that have similar periods in the study area. Schlink et al. (1997) have studied the evolution of smog episodes in Leipzig, and although 3–5 day periods apparently occurred, the spectrum was time dependent and thus non-periodic. In our case, the spectra are essentially time invariant, since sub-series extracted from the complete series give essentially the same periodic information. Thus, it is possible that the short term (high frequency) components observed in all series analysed in this work are due to synoptic weather variations. This is explored below.

4.2. Correlation with synoptic variables

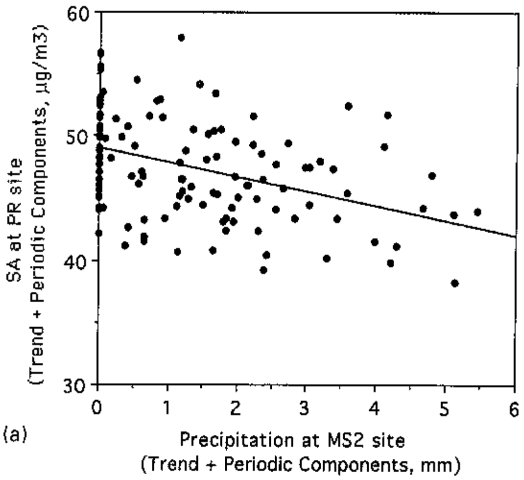
Some information was available from two meteorological stations (MS1 and MS2), located, respectively, near the PR site and south of the river Douro (Fig. 1), the latter in an elevated hill. For MS1, the relative humidity (measured at 9 a.m.), minimum and maximum daily temperatures and precipitation (also measured at 9 a.m.) were analysed for the period of four months, from 1 April to 31 July 1989. For MS2, in addition to the above meteorological variables, the mean wind speed and direction were also available for the same time period. This time period corresponds to a subset for both the PR and MT series (Table 1). As usually done, both relative humidity and precipitation were referred to the previous 24 h (Simmonds and Keay, 1997). Although the meteorological series are rather short, this should be no problem since we are analysing only high-frequency (2–7 days) components.

The meteorological series were analysed with the SATSA program, to detect any underlying periodicities.

For the MS1 site, the only significant periodic components were detected for minimum (2.93 days) and maximum (4.53 days) daily temperatures, respectively, accounting for 1.2 and 1.7% of the total variances. For the MS2 site, periodicities were detected for maximum daily temperatures (4.52 days), precipitation (3.17 and 2.70 days) and mean wind speed (4.36 days), respectively, accounting for 2.2, 13.1 and 4.8% of the corresponding variances. The data shows a very high correlation coefficient between maximum daily temperatures at both sites ($r = 0.990$) but a smaller correlation for minimum daily temperatures ($r = 0.911$), and Fig. 4 shows that the trend and periodic components obtained for maximum daily temperatures at both sites are very similar.

Since the time period for the meteorological series corresponds to a subset for both the PR and MT series, the trends and periodic components obtained for the SA series in these sampling sites were compared with the trends and periodic components obtained in the meteorological series. For the MT site, there is no correlation with any of the meteorological variables (minimum and maximum daily temperatures at the MS1 site and mean wind speed, precipitation and maximum daily temperature at the MS2 site). This may be due to the unrepresentativeness of the meteorological data for this urban sampling site, but further work is necessary to explain this behaviour.

However, for the PR site, Fig. 5a, b show that there are significant correlations (at the 5% level of significance) with precipitation and maximum daily temperatures. Obviously, since the temperatures are highly correlated (Fig. 4), similar correlations occur at the PR site with minimum daily temperatures at the MS1 site or with maximum temperatures at the MS2 site. Since the precipitation data corresponds to the MS2 site, which is far away from the PR sampling site, it may be argued that it would be more correct to use the precipitation data for the MS1 site. However, as stated above, no periodicities could be detected for this site and this meteorological variable, and no decomposition could be performed. It is interesting to note that trend and periodic components of pollution levels tend to decrease as precipitation increases, as expected, and that they tend to increase with both minimum and maximum daily temperatures. However, since for this data precipitation is uncorrelated with both maximum or minimum daily temperatures, these meteorological variables seem to affect independently the pollution levels.



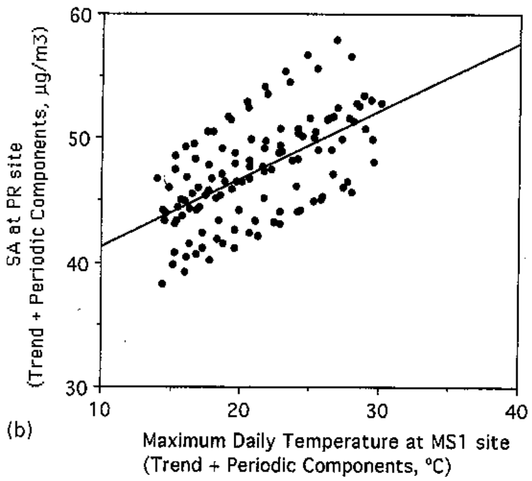


Fig. 5. a. Correlation between trend and periodic components for SA values at the PR site with maximum daily temperature at the MS1 site (1 April–31 July 1989). b. Correlation between trend and periodic components for SA values at the PR site with maximum daily temperature at the MS2 site (1 April–31 July 1989).

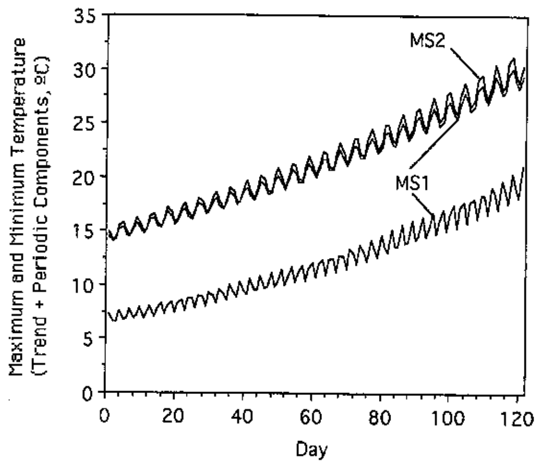


Fig. 4. Trend and periodic components for minimum and maximum daily mean temperatures at the MS1 and MS2 meteorological stations (1 April–31 July 1989).

5. Conclusions

Time-series analysis of air pollution concentrations, both for strong acidity (SA) and black smoke (BS), at three Oporto sampling sites (PR, MT and ML) show that these are mostly random. The application of a sequential decomposition model to air quality data is an adequate method for the identification of the underlying periodicities, despite the low signal-to-noise ratios. High-frequency ($\approx$2–4 day) periodic components and a weekly pattern could be identified in all series, despite being masked by the noisy signals. The periodic components are more important in the BS series, accounting for 10–20% of the observed variance, against 5% in the SA series, which probably reflects that BS values are more representative of pollution originated near the sampling site than SA values. The 7-day period observed for SA concentrations may be related with a weekly pattern of industrial activity, mainly associated with working discontinuous cycles. The behaviour observed for BS concentrations is most probably related to traffic intensity. The application of a structural model based on a sequential decomposition scheme allowed the identification of those cyclical components. However, notwithstanding SATSA being a quite powerful tool, no other cycles than those already known (Benarie, 1980) could be positively recognised in urban air pollution.

The $\approx$2–4 day periods observed for all series may be related to synoptic weather variations with a similar period. The SATSA model was applied to the decomposition of 4-month daily series of precipitation, minimum and maximum temperatures and mean wind speed, obtained at two Oporto meteorological stations. High-frequency ($\approx$2.7–4.5 days) periodic components were observed for maximum and minimum daily temperatures at one station and for maximum daily temperature, precipitation and mean wind speed for the other station. These values were compared with those observed with SA levels (composite trend and periodic components) at both the PR and MT sites, for the same interval of time. Significant positive correlations were obtained with maximum or minimum temperatures, and a negative correlation with precipitation, at the PR site. No significant correlations were obtained at the MT site, which could be due to unrepresentativeness of the meteorological data for this urban sampling site.

Acknowledgements

Support of this work was provided by JNICT, Ministério do Ambiente, AILA and Faculdade de Engenharia da Universidade do Porto.

References

Alvim Ferraz, M.C., Faria Ferraz, M.C., 1988. Evolution of air pollution in Oporto Area. Environmental Monitoring and Assessment 11, 43.

Alvim Ferraz, M.C., Faria Ferraz, M.C., Ferraz, J.M., Fonseca, P.I., 1988. Air pollution and lung diseases in Oporto Area. Environmental Monitoring and Assessment 11, 183.

Alvim Ferraz, M.C., Alves, C.A., 1996. Qualité de l'air dans la région de Porto – analyse comparative. Pollution Atmosphérique Janviers–Mars, 90.

Bartnicki, J., 1989. A simple filtering procedure for removing negative values from numerical solutions of the advection equation. Environmental Software 4(4), 187–201.

Benarie, M.M., 1980. Urban Air Pollution Modelling. Macmillan London and Basingstoke, pp. 228–251.

Benarie, M.M., 1987. The limits of air pollution modelling. Atmospheric Environment 21(1), 1–5.

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis – Forecasting and Control. Holden-Day, San Francisco.

Brigham, E.O., 1974. The Fast Fourier Transform. Prentice-Hall, Englewood Cliffs, NJ.

Chatfield, C., 1989. The Analysis of Time Series – An Introduction. Chapman & Hall, London.

Chock, D.P., Terrl, T.R., Levitt, S.B., 1975. Time series analysis of Riverside, California air quality data. Atmospheric Environments 9, 978–989.

Crabtree, B.F., Ray, S.C., Schmidt. P.M., O'Connor, P.J., Schmidt, D.D., 1990. The individual over time — time series applications in health care research. Journal of Clinical Epidemiology 43(3), 241–260.

Dejak, C., Franco, D., Pastres, R., Pecenik, G., 1990. Irregular environmental historical series — software for statistical and periodic analyses. In: Zannetti, P. (Ed.), Computer Techniques in Environmental Studies III. Computational Mechanics Publications/Springer Berlin, pp. 489–500.

Derwent, R.G., Simmonds, P.G., Seuring, S., Dimmer, C., 1998. Observation and interpretation of the seasonal cycles in the surface concentrations of ozone and carbon monoxide at Mace Head, Ireland, from 1990 to 1994. Atmospheric Environment 32(2), 145–157.

Fuller, Wayne A., 1976. Introduction to Statistical Time Series. John Wiley & Sons, Inc., USA.

Gordon, S.P., Gordon, F.S., 1994. Contemporary Statistics — A Computer Approach. McGraw-Hill, New York.

Hoff, J.C., 1983. A Practical Guide to Box–Jenkins Forecasting. Wadsworth, Balmont, CA.

Schwartz, J., Marcus, A., 1990. Mortality and air pollution in London: a time series analysis. American Journal of Epidemiology 131, 185.

Kendall, Sir M., Ord, J.K., 1990. Time Series, 3rd ed. Arnold, Paris.

Little, J.N., Shure, L., 1993. Signal Processing Toolbox for Use With MATLAB, The MathWorks, Inc., 1–37: 1–39.

Milionis, A.E., Davies, T.D., 1994a. Regression and stochastic models for air pollution — I. Review, comments and suggestions. Atmospheric Environment 28(17), 2801–2810.

Milionis, A.E., Davies, T.D., 1994b. Regression and stochastic models for air pollution — II. Application of stochastic models to examine the links between ground-level smoke concentration and temperature inversions. Atmospheric Environment 28(17), 2811–2822.

Minors, D.S., Waterhouse, J.M., 1988. Mathematical and statistical analysis of circadian rhythms. Psychoeuroendoc. 13, 443–464.

Montgomery, D.C., Johnson, L.A., Gardiner, J.S., 1990. Forecasting and Time Series Analysis, 2nd ed. McGraw-Hill, New York.

Murteira, B.J.F., Muller, D.A., Turkman, K.F., 1993. Análise de Sucessões Cronológicas. McGraw-Hill, Portugal.

Nelson, H.L., Granger, C.W.J., 1979. Experience with using the Box–Cox transformation when forecasting economic time series. Journal of Econometrics 10, 57–69.

Oppenheim, A.V., Shafer, R.W., 1975. Digital Signal Processing. Prentice-Hall, Englewood Cliffs, NJ.

Payne, J.P., 1988. Introduction to Simulation — Programming Techniques and Methods of Analysis. McGraw-Hill International Editions, Computer Science Series.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1986. Numerical Recipes — The Art of Scientific Computing. Cambridge University Press, New York.

Rich, L.G., 1973. Environmental Systems Engineering. McGraw-Hill, New York.

Robinson, E.A., Silvia, M.T., 1979. Digital Foundations of Time Series Analysis — Vol. 1. The Box–Jenkins Approach. Holden-Day, San Francisco, CA.

Salcedo, R.L.R., 1990. Gaussian simulation of SO_2 dispersion emitted from the Petrogal petrochemical complex in the Oporto Metropolitan Area (in Portuguese). Internal Report, FEUP.

Salcedo, R.L.R., Dias, D.C., 1992. A sequential approach to time-series analysis. In: Computer Techniques in Environmental Studies. Computational Mechanics Publication, Elsevier Amsterdam, p. 295.

Schlink, U., Herbarth, O., Tetzlaff, G., 1997. A component time-series model for SO_2 data: forecasting, interpretation and modification. Atmospheric Environment 31(9), 1285–1295.

Schwartz, J., Marcus, A., 1990. Mortality and air pollution in London: a time series analysis. American Journal of Epidemiology 131, 185–194.

Simmonds, I., Keay, K., 1997. Weekly cycle of meteorological variations in Melbourne and the role of pollution and anthropogenic heat release. Atmospheric Environment 31(11), 1589–1603.

Simpson, R.W., Layton, A.P., 1983. Forecasting peak ozone levels. Atmospheric Environment 17, 1649–1654.

Slutzky, E., 1937. The summation of random causes as the sources of cyclic processes (original in Russian, 1927). English trans. Econometrica 5, 105–146.

Tilley, T., McBean, G.A., 1973. An application of spectrum analysis to synoptic-pollution data. Atmospheric Environment 7, 793–801.

Trivikrama, S.R., Samson, P.I., Pedadda, A.R., 1976. Spectral analysis approach to the dynamics of air pollutants. Atmospheric Environmnt 10, 375–379.

Yule, G.U., 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. Philosophical Transactions A 226, 267–298.

Zannetti, P., 1990. Air Pollution Modeling — Theories, Computational Methods and Available Software. Computational Mechanics Publications.