Abstract

Stochastic models that estimate the ground-level ozone concentrations in air at an urban and rural sampling points in South-eastern Spain have been developed. Studies of temporal series of data, spectral analyses of temporal series and ARIMA models have been used. The ARIMA model $(1,0,0) \cdot (1,0,1)_{24}$ satisfactorily predicts hourly ozone concentrations in the urban area. The ARIMA $(2,1,1) \cdot (0,1,1)_{24}$ has been developed for the rural area. In both sampling points, predictions of hourly ozone concentrations agree reasonably well with measured values. However, the prediction of hourly ozone concentrations in the rural point appears to be better than that of the urban point. The performance of ARIMA models suggests that this kind of modelling can be suitable for ozone concentrations forecasting.

Keywords: Urban ozone; Rural ozone; Stochastic models

Contents

1. Introduction ................................................................. 1379
2. Description of the study sites and data collection ........................................... 1380
3. Research and methodology .................................................. 1381
   3.1. Spectral analysis of temporal series (TS) ........................................ 1381
   3.2. ARIMA models ......................................................... 1383
       3.2.1. ARIMA model building in the urban area ................................. 1384
       3.2.2. ARIMA model building for the rural area ................................. 1386
4. Conclusions ............................................................... 1388
   References ................................................................ 1389

1. Introduction

The concentration of ozone in the troposphere is of great interest because of its negative influence on human
health, vegetation and materials. Ground-level ozone is primarily produced from its precursors of NOx and volatile organic compounds (VOC) via complex photochemical reactions in sunlight. Accumulation of ozone near the ground is influenced by physical and chemical processes and by meteorological conditions (NCR, 1991). Atmospheric pollution levels exhibit complex variability over a wide range of spatial and temporal scales with adverse effects on the environment. Often, spatio-temporal variability cannot be accurately represented via physically based (mechanistic) models, e.g. mathematical models of diffusion and transport of pollutants, due to insufficient knowledge of input parameters. The current state of progress in measurement analyses and modelling of ozone precursors, photochemical behaviour and transport processes has been recently reviewed by Sillman (1999), Blanchard (2000), Hidy (2000), Russell and Dennis (2000), Trainer et al. (2000), Ibáñez-Berastegui et al. (2001), Vingarzan and Taylor (2003). Although ozone chemistry has been extensively investigated in many chamber experiments and in photochemical modelling studies, there are still significant difficulties in accurately predicting ambient ozone levels, as well as its spatial distribution, behaviour and associated trends. To track and predict ozone, one must create an understanding of not only ozone itself but also the conditions that contribute to its formation. It is necessary to apply models that describe and help to understand the complex relationships between ozone concentrations and the many variables that cause or hinder ozone production. Photochemical models are often employed to predict hourly ozone variations and thus help to establish cost-effective means of reducing ambient ozone to control, in particular, the emissions of VOC and NOx, from various sources. Because the complex series of reactions are driven by temperature and sunlight, ozone formation varies hourly, daily and seasonally. In the previous paper (Dueñas et al., 2002) an investigation of the importance of meteorology in surface ozone concentrations is presented and linear regression methods for predicting ozone concentrations were used.

In this paper we develop several types of stochastic models to forecast the hourly ozone concentrations in two different sampling points. Predictions turn out to be more interesting because when the ozone air concentrations exceed some levels it is dangerous for human health, vegetation and materials. Stochastic models provide a framework for routing uncertainty into predictions and are being increasingly used for analysing atmospheric pollution levels (Kyriakidis and Journel, 2001). The AutoRegressive Integrated Moving-Average (ARIMA) models require studies of temporal series and spectral analysis of time series of data in both sampling points. The highest ozone concentrations are mostly recorded during the diurnal stage in summer both in urban and rural sampling points. Ozone episodes tended to occur on a warm day with sufficient sunlight, low wind and high surface pressure. Notably, most studies of ozone have examined events during summers in mid-latitudes regions where weather is characterised as “sunny, sultry and stagnant” (Hubbard and Cobourn, 1999). For the urban site the average monthly concentrations throughout the year start to increase in March reaching their maximum values in July. However, in the rural area, the monthly variations are smaller reaching their maximum value in June. The rural concentrations are higher than the urban ones. Summer is the season when there are similar concentrations in both sampling sites. The ozone threshold of the European Union Directive for damages to human health (120 µg m\(^{-3}\), 8 h average) is exceeded systematically for at least four months of the year and the one for the information to the population (180 µg m\(^{-3}\), hourly average) can also be exceeded frequently from April to August. The ozone threshold of the European Union Directive for damages to vegetation (65 µg m\(^{-3}\) as a 24 h average) is exceeded for a 96% of measurements in rural area and a 63% in urban area in summer.

ARIMA models have been developed and validated for both sampling sites. The results are satisfactory though the prediction of hourly ozone concentrations in the rural point is better than that of the urban point.

2. Description of the study sites and data collection

Measurements were carried out at two (urban and rural) sites in Málaga. Details of the sampling points are described in Dueñas et al. (2004). Málaga is the major coastal city of Andalucía region in South Spain and it is a touristic city with a population of 600000. Being Málaga a touristic place, the air quality in the city and its neighbouring villages have been very important. This Spanish city on the Mediterranean coast is distinguished by its mild, temperate and warm climate with low rainfall and approximately 320 days of sun a year. The orographic features play an important role in the interpretation and understanding of ozone behaviour. The coast is backed by a series of mountains that have to be crossed to reach the inland valleys. Due to the influence of the local orography there are a prevailing SE and NW winds (Ortega and Sánchez, 1976). The SE and NW winds occur as a result of the sea-land and land-sea breezes, respectively. The urban site (4°28’8” W; 36°43’40” N) was located on the Faculty of Sciences building, University of Málaga, in the north-eastern part of the city. It was located approximately 5 km from the coastline, near the airport and surrounded by roads with dense traffic. The rural site (4°3’20” W; 36°42’31” N) was located at the La Mayora experimental station in the village of Algarrobo. This
site was 45 km from Málaga and approximately 2 km from the coastline. Although this area has recently experienced a demographic increase, it has always been characterised as an agricultural area especially dedicated to crops being cultivated under plastic. The experimental station covers 51 Has with almost 20,000 m² of greenhouses. The exclusive subtropical climate of this coastal area favours the production of vegetable crops under plastic greenhouses (tomato, melon, watermelon, etc.) and subtropical fruit trees (avocado, custard apple, mango, etc.). These species have a high productive potential and an enormous social and economical importance in this part of the country.

Ozone concentrations were measured from November 1996 to November 1997 in an urban area. The same measurements were carried out in a rural area from February 1998 to October 1999. Ozone levels were continuously monitored using Dasibi Environmental Corporation instrumentation (Dasibi 1008-RS), an ozone monitor based on the absorption of ultraviolet radiation by O₃ at 254 nm as the principle of measurement. This instrument has a limit of detection of 1 ppbv. The analyser has an internal ozone generator and is completely automated by means of the ACR-STACK-ON interface incorporated into the Dasibi 1008-RS. During the sampling period the ozone zero was checked every week measuring a fixed ozone concentration produced by the ozone internal generator. The instrument was continuously operating every day (for 24 h). The interval of each measurement was 2 min and data have been averaged into hourly periods. Air samples were collected through Teflon inlet tubes. The height of the air intake was 2 m above the ground. The measured ozone concentration was represented in units of ppbv and recorded on a strip chart and later with a digital data logger. The instrument was calibrated in the factory by using a stabilised ozone source scaled by a long path UV absorption instrument and periodically compared with the standard ozone calibrator.

3. Research and methodology

The aim of this work has been to develop mathematical models able to describe and subsequently forecast ozone levels. In general, forecast models can be classified as deterministic, empirical and stochastic. Deterministic models are based on the solution of the advection-diffusion equation and allow the study of the spatial and temporal distribution of pollutants. These models have some main disadvantages such as the need to use a high number of meteorological input parameters and the need to have an inventory of pollutant sources. Empirical models are based on the analysis of correlation between concentration levels and the numerous meteorological variables influencing them. The selection of these variables is made empirically and they are included in these models to obtain the best results. Finally stochastic forecast models are based on the analysis of the temporal series that are subject to this study. These models can be univariate or multivariate. The first ones only use the series itself as information source for the design of the model. The second ones, besides the series, can include exogenous variables which give additional information on the series variability. In this work, we develop univariate stochastic models both for the urban and rural area. Our objective has been designing models able to forecast ozone concentration at a certain hour according to preceding data, taking into consideration that the models applied were not designed to forecast episodic situations but to know its behaviour and assess its temporal evolution.

3.1. Spectral analysis of temporal series (TS)

A temporal series (TS) can be defined as the set of observations of a simple variable that has been measured at regular time intervals (Box et al., 1994). A temporal series has memory when every value tends to be near its preceding value, that is to say, when it shows a marked positive correlation. First we analysed the hourly ozone concentrations measured in the urban area. The analysis of temporal series requires having consecutive information about observations of full days. We separated five temporal series, one series for each of the seasons in which data were grouped, except for the summer season, in which two series were taken. Summer period is of great interest in forecasting studies as ozone concentrations are the highest. For these series to be indicative of a certain season, they must satisfy the comparison of mean values, test for comparison of medians by using the Mann–Whitney (Wilcoxon) statistic and the Kolmogorov–Smirnov test to compare distributions (Gondar Nores, 1998). The test for the comparison of mean values constructs confidence intervals or bounds for every mean and for the differences between them. The Mann–Whitney (Wilcoxon) test compares the value of the medians of two samples. It consists in combining the two samples, arranging data in an upward order and comparing the average levels of the samples of combined data. The Kolmogorov–Smirnov test is used to check that both samples come from the same distribution. This test is performed by computing the maximum distance between the cumulative distributions of both samples. The selected series, except for that of autumn, satisfactorily pass the aforementioned tests with a confidence level of 95% or higher.

Once the series were selected, a spectral analysis was undergone. This study was carried out to analyse the statistical significance of periodicities in the series. To study the temporal behaviour, the harmonic components of the data sets were analysed. We obtained a
periodogram based on a Fourier analysis which assumes that the data time series of hourly ozone concentrations were formed by the superposition of sinusoidal components of different frequencies. The Fourier transform quantitatively shows the contribution of every harmonic to the total variance of the data series.

The periodograms of the series representing the winter, spring, summer and autumn seasons for the urban site are shown in Fig. 1. From the spectral analysis performed independently on each season, a 12-h cycle was clearly identified as the most dominant periodic component for the winter and autumn seasons. The observation of such figure allows us to conclude that for the winter and autumn seasons there is a clear cyclic behaviour with a 12-h period. This cycle can be attributed to the hours of sunshine during these seasons. However, the cycle of the spring season is not clearly defined as there are the 0.0416 and 0.0833 frequencies representing the 24-h and 12-h cycles, respectively. The summer season shows a clean periodogram with a single significant peak for the 24-h period. This behaviour coincides with the results obtained by several researchers (Zurita and Castro, 1983; Danalatos and Glavas, 1996; Sebald et al., 2000). In the summer season, the cycle can be attributed to an increase in the number of the hours of sunshine and to a higher solar intensity which favour the photochemical formation of ozone.

Following the same methodology used in the urban area, several temporal series have been selected to proceed to an analysis similar to that carried out in the urban area, ending with the design of an ARIMA forecast model for the most outstanding series. These series are indicative of a certain season if they satisfy the mean values comparison tests, comparison of means by using the Mann–Whitney (Wilcoxon) test and the Kolmogorov–Smirnov test to compare distributions. The studied series were: winter, spring, autumn, summer-1 and summer-2. Out of the five series that were analysed, only that

Fig. 1. Periodograms of the series representing the four seasons in the urban area. The ordinate is equal to the square amplitudes of the sine functions.
of the autumn season did not pass these tests. We carried out a spectral analysis aiming at analysing the statistical significance of periodicities (Fig. 2). It can be concluded from the examination of this figure that the daily cycle is representative for all seasons and the 12-h cycle only seems to influence the winter period, with a contribution much lower than that observed in the urban area.

3.2. ARIMA models

AutoRegressive Integrated Moving-Average (ARIMA) models have been studied extensively by Box and Jenkins (1976). These authors have effectively put together in a comprehensive manner the relevant information required to understand and use univariate time series ARIMA models. Our approach in this section includes two aspects: (1) Use of a simulation program (ARIMA) to generate time-series data, according to a specified ARIMA model. (2) The simulated data from special ARIMA model are analysed to see how closely the empirical properties of the time series match the known theoretical properties.

The general ARIMA model is traditionally described as $ARIMA(p,d,q)$, where ‘p’ is the order of autorregression, ‘d’ is the differentiation degree and ‘q’ is the order of the moving average involved. In the case of seasonal models, the notation is $ARIMA(p,d,q) \times (p,d,q)_T$, where $T$ represents the series periodicity. There is a model building process, described by Box and Jenkins (1976), allowing us to design the best possible model for a TS. This method consists of three main stages (Makridakis et al., 1983):

Identification → Estimation → Diagnostic checking (Validation and Forecast).

These stages successively recur to reach a model yielding satisfactory values. The identification and estimation stages are carried out by means of the analysis of autocorrelation functions (FAC) and the partial

---

**Fig. 2.** Periodograms of the series representing the four seasons in the rural area. The ordinate is equal to the square amplitudes of the sine functions.
autocorrelation function (FACP) of the temporal series. Firstly, the ARIMA methodology requires to determine whether the TS is stationary or not. This was carried out by means of a sequential graph. The validation stage consists of checking that the generated model matches the series generating it. When the model has passed this stage, it can be used to predict future values of the variable. An ideal model in the validation process should have the following characteristics: (i) The residuals of the estimated model approach the behaviour of the white noise. (ii) The model should be stationary and invertible. (iii) The coefficients of the model should be statistically significant and have a low autocorrelation. (iv) The coefficients of the model should be able to represent the series by themselves. (v) The adjustment degree should be high as compared to that of other alternative models.

To meet these conditions, the FAC and FACP of the residual series were analysed. The autocorrelation coefficients of these functions must be small and only one or two high correlation coefficients could go beyond the 95% confidence level at random but this cannot happen in the first lags (Uriel, 1995).

To assess the forecast ability of a model, we should check that it reproduces the real values, both qualitatively and quantitatively. To do that, we should carry out an in-depth statistical analysis by means of graphical representations of the observed and calculated temporal series, the distribution of error frequencies and correlation coefficients between the observed and forecasted values. The Environmental Protection Agency (EPA) and the American Meteorological Society (AMS) establish that every model should be accurate when it comes to forecasting maximum values, both in moment and in value; absence of systematic skewness, absence of total errors, temporal correlations of the forecasted values as opposed to those measured and the spatial alignment (Fernández Nieto, 1995).

3.2.1. ARIMA model building in the urban area

The mentioned methodology, which is used to find an ARIMA prediction model for adjusting the series values, can be applied to ozone concentrations. In this point two data series have been selected in the urban area in order to develop a prediction model. The data for summer-1 \( n = 552, \text{days} = 23 \) from this site were used to develop the prediction model and the data for summer-2 \( n = 528, \text{days} = 22 \) from the same site were used to test the forecast potential of the model. The first step consists in depicting the series subject of study and identifying the possible anomalous values, trends, cycles and seasonality. The behaviour of the selected summer-2 series can be observed in Fig. 3a. In the FAC (Fig. 3b), it can be observed how autocorrelation coefficients exponentially decrease from the first lag and this behaviour recurs every 24 periods, with a small negative autocorrelation towards lag 12. This shows the presence of a 24-h cycle, corroborated by the periodogram of the series.

The FACP correlogram of the summer-1 series can be seen in Fig. 3c. There are two significant peaks in lags 1 and 2 and certain fluctuations and a mixture of sinusoidal oscillations near lag 24 are also observed. As a first approximation, we could consider with the information of FAC and FACP an AR model of order 1 or 2 for the non-seasonality series and a moving average process for the seasonal one. After trying several combinations for parameters \( p \) and \( q \), once the possibility to differentiate the series was dismissed, the ARIMA \((1,0,0)\times(1,0,1)_{24}\) model was identified as suitable. The \(p\)-values of the estimated coefficients AR (1), SAR (1), SMA (1) were statistically valid. Several statistics were used to check the adequacy of the tentatively identified model. These statistics were the root mean square error, the mean absolute error, the mean absolute percentage error, the mean error and the mean percentage error. All of them were statistically significant.

For the forecasting ability of the ARIMA model identified, we have taken into consideration a 120-h validation period. Several statistical tests were carried out to compare them with those of the estimation period (see Table 1). Since there are no significant differences between both periods, we should consider that the estimated parameters are valid. This behaviour is corroborated by means of the Box–Pierce test for the first 48 autocorrelations. A \(p\)-value of 0.15 was obtained allowing us to accept the hypothesis stating that residuals are randomly distributed. This behaviour and by means of the Box–Pierce test is inferred from the analyses of residuals of autocorrelation and partial autocorrelation functions of the residual series summer-2 that shows the absence of significant peaks going beyond the bound of the 95% confidence level.

We concluded the validation stage representing the real series and the forecasted series. As can be appreciated in Fig. 4a, the model satisfactorily fits the real values, being. However, some deficiencies were observed. There were some minimum values in this model where forecasts differ from the real value though the mean error of the forecast is 13.03%. Fig. 4b shows the frequency histogram corresponding to the distribution of error between the real value and that forecasted by the model. We can observe that the distribution is Gaussian. The highest frequencies belong to the lowest errors, decreasing as error percentages increase.

To put an end to modelling procedure and to check that it was suitable for the forecast in the considered season, we applied the ARIMA \((1,0,0)\times(1,0,1)_{24}\) model to the summer-1 series. In the representation of real values versus those forecasted by this model, we can observe that there is a good agreement between both variables.
The analysis of linear regression between real values and those forecasted by this model produced a satisfactory result ($slope = 0.97 \pm 0.02; \ ordinate = 2.9 \pm 1.5$, regression coefficient = 0.934).

The study of the generated residual series also passed the statistical tests. Therefore we can conclude that it is in fact a random series and does not contain basic information of the generating series. Once the different stages for the building of an ARIMA model have been surpassed, we can conclude that the model proposed for the urban summer season is valid both from the point of view of diagnosis and forecasting.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Estimate</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Urban area</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE (mean absolute error)</td>
<td>11.34</td>
<td>10.62</td>
</tr>
<tr>
<td>ME (mean error)</td>
<td>0.73</td>
<td>-2.37</td>
</tr>
<tr>
<td>MPE (mean percentage error)</td>
<td>-7.11</td>
<td>-11.33</td>
</tr>
<tr>
<td><strong>Rural area</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE (mean absolute error)</td>
<td>5.909</td>
<td>5.960</td>
</tr>
<tr>
<td>ME (mean error)</td>
<td>0.249</td>
<td>-0.135</td>
</tr>
<tr>
<td>MPE (mean percentage error)</td>
<td>-0.556</td>
<td>-0.792</td>
</tr>
</tbody>
</table>

(Fig. 4c). The analysis of linear regression between real values and those forecasted by this model produced a satisfactory result ($slope = 0.97 \pm 0.02; \ ordinate = 2.9 \pm 1.5$, regression coefficient = 0.934).

The study of the generated residual series also passed the statistical tests. Therefore we can conclude that it is in fact a random series and does not contain basic information of the generating series. Once the different stages for the building of an ARIMA model have been surpassed, we can conclude that the model proposed for the urban summer season is valid both from the point of view of diagnosis and forecasting.
3.2.2. ARIMA model building for the rural area

Following the same methodology used at the urban area, we also used the summer season to develop the ARIMA methodology in order to obtain a forecasting model useful to represent the behavior of ozone concentration in the rural area. On the basis of the summer-1 series \((n = 1392, \text{days} = 58)\), we designed a stochastic univariate model that was validated and used for prognosis purposes in the summer-2 series \((n = 2064, \text{days} = 86)\). The sequential graph corresponding to the rural summer-1 series is shown in Fig. 5a and the FAC and FACP in Fig. 5b and c, respectively. The high number of lags in FAC, seven to be precise, far beyond the confidence limit, shows a certain lack of stationarity in the summer-1 series. In the FACP there are numerous peaks going beyond the band of the 95% confidence level. It is therefore necessary to take first-order differences, both in the stationary and in the non-stationary time series, to reach the unavoidable premise of stationarity. The joint analysis of these new FAC and FACP functions leads us to suggest the ARIMA stationary model as a suitable one, having the parameters \(ARIMA(2,1,1) \times (0,1,1)_24\). The \(p\)-values of the estimated coefficients AR (1), AR (2), MA(1) and SMA (1) were statistically valid. Additionally, the statistics used to check the adequacy of the tentatively identified model also were statistically significant. The statistics used were the same as those as the urban area.

We proceed then to compare the different error values for the estimate and validation period of the same
Taking into consideration the fact that we have selected a validation period of 260 data from the same series. In Table 1 we can see that there are no great differences between the two validation periods and the values of the different parameters are satisfactory.

We put an end to the validation stage with the representation of the real and the forecasted series to analyse their agreement. As can be observed in Fig. 6a, the model satisfactorily fits the real values since the mean difference between the real value and the forecasted one is below 5%. Fig. 6b shows the frequency histogram corresponding to the error distribution between the real value and that forecasted by the model. We can observe that the distribution is Gaussian and that the highest frequencies belong to the lowest error, decreasing as error percentages increase.

This model met all the requirements to be considered as suitable, then the forecast stage used the summer-2 series for this purpose. Fig. 6c shows the values forecasted by the model versus the real values of ozone concentration for the considered period. The analysis of linear regression between the real values and those forecasted by the model produced a satisfactory result ($slope = 0.98 \pm 0.01$; $ordinate = 1.5 \pm 0.6$, regression coefficient = 0.962).

Once the different stages for the building of an ARIMA model satisfactorily passed, we can conclude that the proposed model for the rural summer season
is valid both from the point of view of diagnosis and prognosis.

4. Conclusions

Stochastic models have been developed in order to forecast the hourly ozone concentrations in two sites (urban and rural) in South-eastern Spain. The ARIMA models have been estimated and validated for the summer season. The period chosen is of interest because it is when harmful maximum ozone concentrations occur in the area.

To validate the models prediction capability, in both sampling points the data set namely summer-2 were used to test the developed model. The resulting predictions were then compared with actual results and statistical numerical measurements were then calculated. The ARIMA model $(1,0,0) \times (1,0,1)_24$ predicts hourly ozone concentrations in the urban area. The ARIMA $(2,1,1) \times (0,1,1)_24$ was developed for the rural area. The forecast performance of each model was evaluated using graphical and statistical comparisons. Predictions of hourly ozone concentrations agree reasonably well with measured values. However, the prediction of hourly ozone concentrations in the rural site appears to be bet-
ter than that of the urban point. Both models are able to forecast ozone concentrations at a certain hour according to preceding data, taking into consideration that the models applied were not designed to forecast episodic situations but to know their behaviour and assess its temporal evolution. There are not exogenous variable in the develop ARIMA models design. The predictions results for the univariate model were satisfactory. However further comparison between site specific models is an important area for further work because such a comparison may aid in developing site characterisations. These characterisations may prove important for extensions of the developed models to new sites around the world and may aid in creating more guidelines for site specific model development.

References